University of Dundee

DOCTOR OF PHILOSOPHY

Explainable Argument Mining

Lawrence, John

Award date:
2021

Explainable Argument Mining

John Richard Lawrence

Doctor of Philosophy
University of Dundee
May 2021

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank, first and foremost, my supervisor Prof. Chris Reed for the support, encouragement and enthusiasm he has shown, not only in this work, but throughout my career at ARG-tech. It could not be more appreciated.

I would also like to thank my second supervisor Dr. Katarzyna Budzynska for encouraging me to pursue a PhD in the first place, and for her kind support throughout.

My thanks go to the members of ARG-tech, both past and present, who have provided invaluable feedback on this work, and, perhaps even more importantly, made the process so very much more enjoyable.

Finally, my thanks go to my friends and family who have been, as always, an endless source of support and encouragement throughout this journey.

Declaration

I, John Lawrence, hereby declare that I am the author of this thesis; that I have consulted all references cited; that I have done all the work recorded by this thesis; and that it has not been previously accepted for a higher degree.

Abstract

As the volume of data we produce continues to grow, manual techniques increasingly struggle to keep up with the pace at which it is being generated, and greater emphasis is being placed on the automatic extraction of meaning from this data. Opinion mining and sentiment analysis provide valuable information on the views expressed in a text, however, they tell us only *what* opinions are being put forth and not *why* people hold the opinions they do. This is the task addressed by argument mining. The majority of argument mining techniques explored to date have focused on applying existing computational linguistic techniques to identify specific facets of the argumentative structure (for example, classifying premise/conclusion or argument/non-argument). The techniques presented in this thesis complement and extend these existing approaches by taking as a starting point the rich heritage of philosophical research in the analysis and understanding of argumentation, and drawing inspiration from the ways in which humans understand the structure of an argument.

The argument mining techniques presented here cover: a study of explicit linguistic expressions of the relationship between statements (e.g. "because", "therefore" or "however"); contextual knowledge in the form of **premise-conclusion topic models** which capture common patterns of statements matching one topic being used to support or attack statements matching another topic; relating **similarity and topical changes** to underlying argumentative structure; **properties of large scale argument networks** such as how central a proposition is to the text, offering a clue to the argumentative structure often intuitively employed by a human annotator, who will naturally connect a range of supporting arguments to a central conclusion; and **argumentation schemes**, common patterns of human reasoning which have been detailed extensively in philosophy and psychology.

Whilst each of these approaches produces reliable results, illuminating a facet of the full argumentative structure, it is in their combination that these techniques find their greatest strength. The final part of the work presented here looks at combining the output from these individual approaches whilst maintaining explainability of where the structure comes from. Allowing us, for example, to say that there is an inference relation between x and y because they form an instance of a particular argument scheme, or between y and z because of the presence of a discourse indicator. By leveraging the strengths of each, this combined explainable approach is shown to achieve an identification of the argumentative structure that is both more detailed and more accurate than existing argument mining techniques when tested on a corpus of debate from the US 2016 Presidential election, and comparable results to state of the art techniques when tested on widely used third-party corpora.

The work presented in this thesis offers two principal contributions, the development of a range of argument mining techniques grounded in argumentation theory, and, the introduction of *Explainable Argument Mining* (XAM).

Chapter 1

Introduction

As research on specific tasks in data mining has matured, it has been picked up commercially and enjoyed rapid success, with, for example, the sentiment analytics market alone estimated to reach approximately $6bn by 2023[1]. Existing techniques are, however, limited in their ability to identify more complex structural relationships between concepts. Although opinion mining and sentiment analysis provide techniques which are proving to be enormously successful in marketing and public relations (where major brands use the techniques to track opinion of both their own and competitor brands amongst existing and potential customer groups), and in financial market prediction (where large-scale aggregation of sentiment can be used to give insight into likely trends), they can only tell us *what* opinions are being expressed and not *why* people hold the opinions they do.

The study of argumentation, and in particular, the analysis of argument structure, aims to address this issue by turning unstructured text into structured argument data and thereby giving an understanding not just of the individual points being made, but of the relationships between them and how they work together to support (or undermine) the overall message. Figure 1.1, for example, shows an analysis of the argumentative structure contained in the following text:

> Trump's speech was poor. The speech was "lacking in policy prescriptions," and its "strident rhetoric masked a lack of depth," said Robert McFarlane, a former national security

[1] https://www.marketresearchfuture.com/reports/sentiment-analytics-market-4304

adviser. However, his popularity in the polls continues to rise, perhaps because of his recently self-declared high IQ.

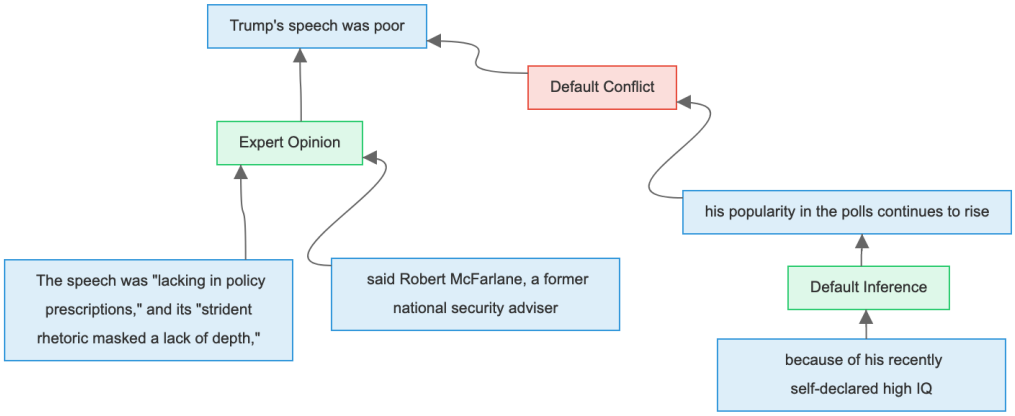


Figure 1.1: An example of analysed argumentative structure

This analysis shows the conflict between the statements that "Trump's speech was poor" and "his popularity in the polls continues to rise", with the former being supported by the statements of former national security adviser Robert McFarlane shown as an instance of argument from expert opinion (where the words of someone knowledgeable in a field are used to support a given claim), and the latter supported by the suggested reason of "his recently self-declared high IQ".

Whilst there is evidence that argument analysis aids comprehension of large volumes of data, the manual extraction of argument structure is a skilled and time consuming process. For example, Robert Horn talking about the argument maps he produced on the debate as to whether computers can think, quotes a student as saying "These maps would have saved me 500 hours of time my first year in graduate school"[2], however Metzinger (1999) notes that over 7,000 hours of work was required in order for Horn and his team to create these maps.

Although attempts have been made to increase the speed of manual argument analysis (Bex et al., 2013), it is clearly impossible to keep up with the rate of data being generated across even a small subset of domains and, as such,

[2] `http://www.stanford.edu/~rhorn/a/topic/phil/artclTchngPhilosphy.html`

attention is increasingly turning to Argument Mining[3](Stede and Schneider, 2018; Lawrence and Reed, 2020), the automatic identification and extraction of argument components and structure.

The field of Argument Mining has been expanding rapidly in recent years with ACL workshops on the topic being held annually, from the first in 2014[4], up to the most recent in 2019[5] which received a record number of 41 submissions. Academic research groups as well as commercial initiatives such as IBM's Project Debater[6] are pushing forward our ability to understand the arguments contained in natural language text.

Due in part to this rapid growth and development, the majority of argument mining techniques explored to date have focused on applying existing computational linguistic techniques to identify specific facets of the argumentative structure (for example, classifying premise/conclusion or argument/non-argument). The techniques presented in this thesis complement and extend these existing approaches by taking as a starting point the rich heritage of philosophical research in the analysis and understanding of argumentation, and drawing inspiration from the ways in which humans understand the structure of an argument.

By virtue of this approach, the work presented here offers two principal contributions: the development of a range of argument mining techniques grounded in argumentation theory (see Section 1.1); and, the introduction of *Explainable Argument Mining* (XAM).

Explainability in Artificial Intelligence (Adadi and Berrada, 2018; Gunning et al., 2019) is a growing concern, with, for example, 67% of business leaders surveyed in the PwC global CEO survey[7] stating that they believe "AI and automation will impact negatively on stakeholder trust levels". Such concerns are especially prevalent in areas dealing with sensitive decisions (such as legal judgements), or involving strict editorial control (such as journalism). XAM addresses these issues, providing results which match state of the art techniques

[3]Sometimes also referred to as argumentation mining

[4]`http://www.uncg.edu/cmp/ArgMining2014/`

[5]`https://argmining19.webis.de/`

[6]`https://www.research.ibm.com/artificial-intelligence/project-debater/`

[7]`https://pwc.to/2pZTNuJ`

for accuracy, whilst also drawing inspiration from the ways in which humans understand the structure of an argument to explain the decisions made in the machine extracted argument structure (for example to say that the system believes there is an inference relation between x and y because they form an instance of a particular argument scheme, or between y and z because of the presence of a discourse indicator). XAM has already been adopted as a key component of the BBC Evidence Toolkit project[8], an online application designed to encourage users to dissect and critically appraise the internal reasoning structure of news reports which uses XAM to automatically analyse user selected news stories. XAM unlocks a wide range of potential future applications, a number of which are explored in Chapter 10.

1.1 Explainable Argument Mining Techniques

The first, and simplest, technique presented here is that of using **discourse indicators** to determine the argumentative connections between adjacent propositions in a piece of text. Discourse indicators are linguistic expressions of the relationship between statements (e.g. "because", "therefore" or "however"), and the identification of such indicators in a text is often the first method of analysing argumentative structure taught to students. The results show that, whilst such indicators can predict with high accuracy the argumentative structure of a text in those cases where they occur, they are very rarely present in real-world arguments. Fewer than a third of the argumentative relations in the corpus studied here are marked by any kind of discourse indicator, meaning that this technique on its own is clearly insufficient. Figure 1.1 shows two examples of this, with the default conflict relation being indicated with "however", and the default support being indicated with "because".

Those discourse indicators that have been shown to have the highest precision can, however, be used to harvest *weakly labelled data* (data which has not been explicitly labelled, but for which certain assumptions can be said to hold) on a given topic, creating a corpus of inference (or conflict) relations

[8] `https://www.bbc.co.uk/taster/pilots/evidence-toolkit-moral-maze`

that are common in that topic area. This captures the idea of an analyst using contextual knowledge of the common arguments on a topic to identify potential support and conflict relations. In this technique, the weakly labelled data is used to create **premise-conclusion topic models** which capture an understanding of an argument being made, not just from the words said, but from an understanding of the broader issues. This weakly labelled data is used to produce a matrix representing the inferential relationships between different aspects of the topic, and from this matrix, we are able to determine inference relations between statements in the original text.

Knowledge of the topics under discussion is also key to following lines of reasoning, relating **similarity and topical changes** to the underlying argumentative structure. This technique starts with the hypotheses that, firstly, the argument structure to be determined can be represented as a tree, and secondly, that this tree is generated depth first. Based on these assumptions we can determine the structure by looking at how similar the topic of each proposition is to its predecessor. If they are similar, then we assume that they are connected and the line of reasoning is being followed. If they are not sufficiently similar, then we first consider whether we are moving back up the tree, and so compare the current proposition to all of those made previously and connect it to the most topically similar previous point. Finally, if the current point is not related to any of those made previously, then it is assumed to be unconnected to the existing structure. Figure 1.1 highlights this technique, with the topic first relating to Trump's speech, before then moving on to look at his popularity in the polls.

Some topics, or even particular propositions may be more central to the argument than others. Again, this offers a clue to the argumentative structure often intuitively employed by a human annotator, who will naturally connect a range of supporting arguments to a central conclusion. This is studied by considering the **properties of large scale argument networks** as a whole, and looking at the complex interactions between their constituent propositions. We investigate metrics for analysing properties of these networks, and present techniques for determining these features directly from natural language text.

We show that there is a strong correlation between these automatically identified features and the argumentative structure contained within the text.

Such patterns in the argumentative structure also exist at a more fine-grained level, and are captured by **argumentation schemes**; patterns of human reasoning which have been detailed extensively in philosophy and psychology. In the final approach presented here, it is demonstrated that the structure of such schemes can provide rich information for the task of automatically identifying complex argumentative structures. By training a range of classifiers to identify the individual proposition types which occur in these schemes, it is possible not only to determine where a scheme is being used, but also the roles played by its component parts. This work extends that already carried out on scheme identification, removing the need for the structure to have already been determined, and providing valuable 'partial' results where some of the components of a scheme instance are correctly identified.

Whilst each of these approaches produces reliable results, illuminating a facet of the full argumentative structure, it is in their combination that these techniques find their greatest strength. The final part of the work presented here looks at combining the output from these individual approaches whilst maintaining explainability of where the structure comes from. It is shown that, by leveraging the strengths of each, it is possible to achieve an identification of the argumentative structure that is both more detailed and more accurate than existing argument mining techniques.

1.2 Outline

The main content of this thesis begins in Chapter 2 with an in-depth review of the argument mining field. This includes: foundational techniques that relate to argument mining (Opinion Mining, Controversy Detection, Citation Mining, and Argumentative Zoning); a detailed look at manual argument analysis providing inspiration for many of the techniques presented in later chapters; a framework for breaking down existing argument mining work into different application areas (Identifying Argument Components, Identifying Clausal

Properties, Identifying Relational Properties); and a thorough exploration of existing work in each of these areas.

In Chapter 3, Argument Data is explored, looking first at available argument corpora, then techniques to automatically generate or extend the volume of argument data available, before finally looking in detail at the corpora used throughout the rest of the work presented here.

Chapters 4-8 then each introduce one of the techniques developed for this work. These are: Discourse Indicators, providing the first study of the prevalence of such indicators and exploring their applicability to the argument mining task; Premise-Conclusion Topic Models, an approach to automatically generate topic models from online data, representing common themes of inference or support, on any given topic; Similarity, looking at a range of methods to determine whether a pair of propositions are in some way similar in what they express, and studying the connections between this similarity and argument structure; Graph Properties, investigating whether properties of large scale argument graphs, such as the centrality of a particular proposition in the graph, can be determined directly from linguistic cues in the text, and then used to determine the argument structure, rather than the other way around; Argument Schemes, presenting a technique for identifying individual scheme components, and using these to both automatically label schemes, as well as give clues as to how they fit in a larger argument structure.

Chapter 9 combines all of these approaches, presenting a rule-based method of combination, which has the advantage of maintaining the explainability inherent in each individual approach. A representation of these results, compliant with the Argument Interchange Format (Chesñevar et al., 2006) and Inference Anchoring Theory (Budzynska and Reed, 2011) is further proposed, showing how the reasons for the decisions made can be viewed as supporting Searle's (Searle, 1969) first preparatory rule for assertion (the speaker has evidence (reasons etc.) for the truth of the proposition being asserted). The results for the rule based combination method are compared to, and shown to outperform, a number of machine learning based methods of combining the same. This shows that not only does such a rule-based approach maintain

explainability, but also does not lose out in performance compared to alternative combination approaches. Finally, the rule based combination method is evaluated against two widely used argumentation corpora (the Argument Annotated Essays corpus (Stab and Gurevych, 2017), and the Argumentative Microtext corpus (Peldszus and Stede, 2016)) and the results compared to existing work on these datasets. This comparison shows that the combined approach can produce comparable results to state of the art techniques developed specifically for use on this data.

Finally, Chapter 10 looks ahead to a number of potential downstream applications of argument mining. These range from applications which rely directly on argument mining algorithms to provide their functionality, to software for visualising and analysing arguments once the argumentative structure has been successfully mined.

1.3 Published Work

The different chapters composing this work are, in many cases, extended and revised versions of published research works. Below is a collected list of the published articles which have formed the foundation of each chapter. Where any of these papers have multiple authors, only work directly contributed by the author of this thesis is included in these chapters. The most relevant texts are marked in **bold**.

- Chapter 2: Literature Review
  - **Lawrence, J. and Reed, C. (2020). Argument mining: A survey. Computational Linguistics, 45(4):765–818**
- Chapter 3: Argument Data
  - Visser, J., Duthie, R., Lawrence, J., and Reed, C. (2018a). Intertextual correspondence for integrating corpora. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC), pages 3511–3517, Miyazaki, Japan. European Language Resources Association (ELRA)

– **Reed, C., Budzynska, K., Duthie, R., Janier, M., Konat, B., Lawrence, J., Pease, A., and Snaith, M. (2017). The argument web: an online ecosystem of tools, systems and services for argumentation. Philosophy & Technology, 30(2):137–160**

– Konat, B., Lawrence, J., Park, J., Budzynska, K., and Reed, C. (2016). A corpus of argument networks: Using graph properties to analyse divisive issues. In Proceedings of the 10th edition of the Language Resources and Evaluation Conference

– Duthie, R., Lawrence, J., Budzynska, K., and Reed, C. (2016b). The CASS technique for evaluating the performance of argument mining. In Proceedings of the 3rd Workshop on Argumentation Mining, pages 40–49, Berlin, Germany. Association for Computational Linguistics

– **Lawrence, J., Janier, M., and Reed, C. (2015). Working with open argument corpora. In Proceedings of the 1st European Conference on Argumentation (ECA 2015), Lisbon. College Publications**

– **Lawrence, J. and Reed, C. (2014). AIFdb corpora. In Parsons, S., Oren, N., Reed, C., and Cerutti, F., editors, Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014), pages 465–466, Pitlochry, Scotland. IOS Press**

– Janier, M., Lawrence, J., and Reed, C. (2014). OVA+: An argument analysis interface. In Parsons, S., Oren, N., Reed, C., and Cerutti, F., editors, Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014), pages 463–464, Pitlochry, Scotland. IOS Press

– Bex, F., Lawrence, J., Snaith, M., and Reed, C. (2013). Implementing the argument web. Communications of the ACM, 56(10):66–73

– **Lawrence, J., Bex, F., Reed, C., and Snaith, M. (2012b).**

AIFdb: Infrastructure for the argument web. In Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012), pages 515–516, Vienna, Austria. IOS Press

  - Bex, F., Gordon, T. F., Lawrence, J., and Reed, C. (2012). Interchanging arguments between Carneades and AIF – Theory and practice. In Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012), pages 390–397, Vienna, Austria. IOS Press

- Chapter 4: Discourse Indicators

  - **Lawrence, J. and Reed, C. (2015). Combining argument mining techniques. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 127–136, Denver, CO. Association for Computational Linguistics**

- Chapter 5: Premise-Conclusion Topic Models

  - **Lawrence, J. and Reed, C. (2017a). Mining argumentative structure from natural language text using automatically generated premise-conclusion topic models. In Proceedings of the 4th Workshop on Argument Mining, pages 39–48, Copenhagen, Denmark. Association for Computational Linguistics**

- Chapter 6: Similarity

  - **Lawrence, J., Park, J., Budzynska, K., Cardie, C., Konat, B., and Reed, C. (2017a). Using argumentative structure to interpret debates in online deliberative democracy and erulemaking. ACM Transactions on Internet Technology (TOIT), 17(3):25**

  - Murdock, J., Allen, C., Borner, K., Light, R., McAlister, S., Ravenscroft, A., Rose, R., Rose, D., Otsuka, J., Bourget, D., Lawrence, J.,

and Reed, C. (2017). Multi-level computational methods for interdisciplinary research in the hathitrust digital library. PLOS ONE, 12(9):1–21

  - Lawrence, J., Reed, C., Allen, C., McAlister, S., and Ravenscroft, A. (2014). Mining arguments from 19th century philosophical texts using topic based modelling. In Proceedings of the First Workshop on Argumentation Mining, pages 79–87, Baltimore, MD. Association for Computational Linguistics

- Chapter 7: Graph Properties

  - **Lawrence, J. and Reed, C. (2017b). Using complex argumentative interactions to reconstruct the argumentative structure of large-scale debates. In Proceedings of the Fourth Workshop on Argumentation Mining, Copenhagen. Association for Computational Linguistics**

- Chapter 8: Agumentation Schemes

  - **Visser, J., Lawrence, J., Reed, C., Wagemans, J., and Walton, D. (2021). Annotating argument schemes. Argumentation, 35:101–139**

  - Visser, J., Lawrence, J., Wagemans, J., and Reed, C. (2018c). Revisiting computational models of argument schemes: Classification, annotation, comparison. In Modgil, S., Budzynska, K., and Lawrence, J., editors, Proceedings of the Seventh International Conference on Computational Models of Argument (COMMA 2018), pages 313–324, Warsaw. IOS Press

  - **Lawrence, J. and Reed, C. (2016). Argument mining using argumentation scheme structures. In Baroni, P., Stede, M., and Gordon, T., editors, Proceedings of the Sixth International Conference on Computational Models of Argument (COMMA 2016), pages 379–390, Potsdam, Germany. IOS Press**

- Chapter 10: Applications of Argument Mining

– **Lawrence, J., Visser, J., and Reed, C. (2018). BBC Moral Maze: Test your argument. In Modgil, S., Budzynska, K., and Lawrence, J., editors, Proceedings of the Seventh International Conference on Computational Models of Argument (COMMA 2018), pages 465–466, Warsaw. IOS Press**

– **Lawrence, J., Snaith, M., Konat, B., Budzynska, K., and Reed, C. (2017b). Debating technology for dialogical argument: Sensemaking, engagement, and analytics. ACM Transactions on Internet Technology (TOIT), 17(3):24:1–24:23**

– Pease, A., Lawrence, J., Budzynska, K., Corneli, J., and Reed, C. (2017). Lakatos-style collaborative mathematics through dialectical, structured and abstract argumentation. Artificial Intelligence, 246:181–219

– Snaith, M., Medellin, R., Lawrence, J., and Reed, C. (2017). Arguers and the Argument Web. In Bex., F., Grasso, F., Green, N., Paglieri, F., and Reed, C., editors, Argument Technologies: Theory, Analysis & Applications, pages 57–72. College Publications

– **Lawrence, J., Duthie, R., Budzysnka, K., and Reed, C. (2016). Argument analytics. In Baroni, P., Stede, M., and Gordon, T., editors, Proceedings of the Sixth International Conference on Computational Models of Argument (COMMA 2016), pages 371–378, Berlin. IOS Press**

– Bex, F., Lawrence, J., and Reed, C. (2014a). Generalising argument dialogue with the dialogue game execution platform. In Parsons, S., Oren, N., Reed, C., and Cerutti, F., editors, Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014), pages 141–152, Pitlochry, Scotland. IOS Press

– Bex, F., Snaith, M., Lawrence, J., and Reed, C. (2014b). Argublogging: An application for the argument web. Web Semantics: Science, Services and Agents on the World Wide Web, 25:9–15

- Pease, A., Budzynska, K., Lawrence, J., and Reed, C. (2014). Lakatos games for mathematical argument. In Parsons, S., Oren, N., Reed, C., and Cerutti, F., editors, Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014), pages 59–66, Pitlochry, Scotland. IOS Press
- **Lawrence, J., Bex, F., and Reed, C. (2012a). Dialogues on the argument web: Mixed initiative argumentation with arvina. In Proceedings of the 4th International Conference on Computational Models of Argument (COMMA 2012), pages 513–514, Vienna, Austria. IOS Press**

• Chapter 11: Conclusion

- Lawrence, J., Visser, J., and Reed, C. (2019a). An online annotation assistant for argument schemes. In Proceedings of the 13th Linguistic Annotation Workshop, pages 100–107, Florence, Italy. Association for Computational Linguistics
- Lawrence, J., Visser, J., Walton, D., and Reed, C. (2019b). A decision tree for annotating argumentation scheme corpora. In 3rd European Conference on Argumentation (ECA 2019), pages 97–114, Groningen, Netherlands
- Visser, J., Duthie, R., Lawrence, J., and Reed, C. (2018b). Intertextual Correspondence for Integrating Corpora. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pages 3511–3517, Miyazaki, Japan. European Language Resources Association (ELRA)
- **Lawrence, J., Visser, J., and Reed, C. (2017c). Harnessing rhetorical figures for argument mining. Argument & Computation, 8(3):289–310**

Chapter 2

Literature Review

In this chapter, we look first, in Section 2.1, at existing work in areas which form the foundation for many of the current approaches to argument mining, including opinion mining, controversy detection, citation mining and argumentative zoning. In Section 2.2 we look at the task of manual argument analysis, considering the steps involved and tools available, as well as the limitations of manually analysing large volumes of text. Section 2.3, provides an overview of the tasks involved in argument mining before giving a comprehensive review of each in Sections 2.4, 2.5 and 2.6.

2.1 Foundational Areas and Techniques

In this section, we look at a range of different areas which constitute precursors to the task of argument mining. Although these areas are somewhat different in their goals and approach, they all offer techniques which at least form a useful starting point for determining argument structure. We do not aim to present a comprehensive review of these techniques in this section, but instead, to highlight their key features and how they relate to the task of argument mining.

In Section 2.1.1, we present an overview of Opinion Mining, focusing specifically on its connection to argument mining. Section 2.1.2 looks at Controversy Detection, an extension of opinion mining which aims to identify topics where opinions are polarised. Citation Mining, covered in Section 2.1.3, looks at

citation instances in scientific writing and attempts to label them with their rhetorical roles in the discourse. Finally, in Section 2.1.4, we look at Argumentative Zoning, where scientific papers are annotated at the sentence level with labels that indicate the rhetorical role of the sentence (criticism or support for previous work, comparison of methods, results or goals, etc.).

2.1.1 Opinion Mining

As the volume of online user-generated content has increased, so too has the availability of a wide range of text offering opinions about different subjects, including product reviews, blog posts and discussion groups. The information contained within this content is valuable not only to individuals, but also to companies looking to research customer opinion. This demand has resulted in a great deal of development in techniques to automatically identify opinions and emotions.

Opinion mining is "the computational study of opinions, sentiments and emotions expressed in text" (Liu, 2010). The terms 'opinion mining' and 'sentiment analysis' are often used interchangeably. Although much of the published work mentioning sentiment analysis focuses on the specific application of classifying reviews by polarity (either positive or negative), Pang and Lee (2008) point out that "many construe the term more broadly to mean the computational treatment of opinion, sentiment, and subjectivity in text".

The link between sentiment, opinion and argumentative structure is described in Hogenboom et al. (2010), where the role that argumentation plays in expressing and promoting an opinion is considered and a framework proposed for incorporating information on argumentation structure into the models for sentiment discovery in financial news. Based on their role in the argumentation structure, text segments are assigned different weights relating to their contribution to the overall sentiment. Conclusions, for example, are hypothesised to be good summaries of the main message in a text and therefore key indicators of sentiment. The interesting point here, from an argument mining perspective, is that this theory could equally be reversed and sentiment be used as an indicator of the argumentative process found in a text. Taking the

example of conclusions, those segments which align with the overall sentiment of the document are more likely to be a conclusion than those which do not.

Many applications of sentiment analysis are carried out at the document level to determine an overall positive or negative sentiment. For example, in Pang et al. (2002) topic-based classification using the two "topics" of positive and negative sentiment is carried out. To perform this task, a range of different machine learning techniques (including Support Vector Machines (Cortes and Vapnik, 1995), Maximum Entropy and Naïve Bayes (Lewis, 1998)) are investigated. Negation tagging is also performed using a technique from (Das and Chen, 2001) whereby the tag `NOT_` is prepended to each of the words between a negation word ("not", "isn't", "didn't", etc.) and the first punctuation mark occurring after the negation word. In terms of relative performance, the support vector machines (SVMs) achieved the best results, with average three-fold cross-validation accuracies over 0.82 for positive/negative sentiment classification, on a corpus of 700 randomly selected positive-sentiment and 700 randomly selected negative-sentiment movie reviews extracted from the Internet Movie Database (IMDb) archive[1].

Shorter spans of text are also considered in Grosse et al. (2012), which looks at microblogging platforms such as Twitter with the aim of mining opinions from individual posts to build an "opinion tree" which can be built recursively by considering arguments associated with incrementally extended queries. Sentiment analysis tools are used to determine the overall sentiment for an initial one word query, which is then extended and the change in overall sentiment recalculated. By following this procedure, it is possible to see where extending the query results in a change of overall sentiment and, as such, to determine those terms which introduce conflict with the previous query. Conflicting elements in an opinion tree are then used to generate a "conflict tree", similar to the dialectical trees (Prakken, 2005) used traditionally in defeasible argumentation (Pollock, 1987).

Opinion mining, however, is not limited to just determining positive and negative views. In Kim and Hovy (2006b) sentences from online news media

[1] `http://reviews.imdb.com/Reviews`

texts are examined to determine the topic and proponent of opinions being expressed. The approach uses semantic role labelling to attach an opinion holder and topic to an opinion-bearing word in each sentence using FrameNet[2] (a lexical database of English, based on manual annotation of how words are used in actual texts). To supplement the FrameNet data, a clustering technique is used to predict the most probable frame for words which FrameNet does not include. This method is split into three subtasks:

1. collection of opinion words and opinion-related frames - 1,860 adjectives and 2,011 verbs classified into positive, negative and neutral. Clustering By Committee (Pantel, 2003) is used to find the closest frame. CBC uses the hypothesis that words that occur in the same context tend to be similar.

2. semantic role labelling for those frames. A Maximum Entropy model is used to classify frame element types (e.g. Stimulus, Degree, Experiencer etc.)

3. mapping of semantic roles to the opinion holder and topic. A manually built mapping table maps Frame Elements to a holder or topic.

Results show an increase from the baseline of 0.30 to 0.67 for verb target words and of 0.38 to 0.70 for adjectives, with the identification of opinion holders giving a higher F-score[3] than topic identification.

Although understanding the sentiment of a document as a whole could be a useful step in extracting the argument structure, the work carried out on sentiment analysis at a finer-grained level perhaps offers greater benefit still. In Wilson et al. (2005), an approach to phrase-level sentiment analysis is presented, using a two-step process: first, applying a machine learning algorithm

[2] `https://framenet.icsi.berkeley.edu/`

[3] F-score refers to the equally weighted harmonic mean of the precision and recall measured for a system. When the system is applied to several sets of data, the micro-average F-score is obtained by first summing up the individual true positives, false positives, and false negatives and then calculating precision and recall using these figures, whereas the macro-average F-score is calculated by averaging the precision and recall of the system on the individual sets (van Rijsbergen, 1979)

to classify a phrase as either neutral or polar (for which an accuracy of 0.76 is reported); and then looking at a variety of features in order to determine the contextual polarity (*positive*, *negative*, *both* or *neutral*) of each polar phrase (with an accuracy of 0.62–0.66 depending on the features used).

In Sobhani et al. (2015), we see an example of extending simple pro and con sentiment analysis, to determine the stance which online comments take towards an article. Each comment is identified as "Strongly For", "For", "Other", "Against", and "Strongly Against" the original article. These stances are then linked more clearly to the argumentative structure by using a topic model to determine what is being discussed in each comment, and classify it to a hierarchical structure of argument topics. This combination of stance and topic hints at possible argumentative relations – for example, comments about the same topic that have opposing stance classifications are likely to be connected by conflict relations, whereas those with similar stance classifications are more likely to connect through support relations.

In Kim and Hovy (2006a), the link between argument mining and opinion mining is clearer still. Instead of looking solely at whether online reviews are positive or negative, a system is developed for extracting the reasons *why* the review is positive or negative. Using reviews from epinions.com, which allows a user to give their review as well as specific positive and negative points, these specific positive and negative phrases were first collected and then the main review searched for sentences which covered most of the words in the phrase. Using this information, sentences were classified as "pro" or "con" with unmatched sentences classified as "neither". Sentences from further reviews were then classified as, first, "pro" and "con" against "neither" followed by classification into "pro" or "con". The best feature selection results in an F-score of 0.71 for reason identification and 0.61 for reason classification.

2.1.2 Controversy Detection

One extension to the field of opinion mining that has particular relevance to argument mining is controversy detection, where the aim is to identify controversial topics and text where conflicting points of view are being presented.

The most clear link between controversy and argument detection can be seen in Boltužić and Šnajder (2015), where argumentative statements are clustered based on their textual similarity, in order to identify prominent arguments in online debates. Controversy detection to date has largely targeted specific domains: (Kittur et al., 2007) for example, looks at the cost of conflict in producing Wikipedia articles, where conflict cost is defined as "excess work in the system that does not directly lead to new article content". Conflict Revision Count (CRC), a measure counting the number of revisions in which the "controversial" tag was applied to the article, is developed and used to train a machine learning model for predicting conflict. Computing the CRC for each revision of every article on Wikipedia resulted in 1,343 articles for which the CRC score was greater than zero (meaning they had at least one "controversial" revision). 272 of these articles were additionally marked as being controversial in their most recent revision. A selection of these 272 articles is then used as training data for an SVM classifier. Features are calculated from the specific page such as the length of the page, how many revisions were carried out, links from other articles, and the number of unique editors. Of these features, the number of revisions carried out is determined to be the most important indicator of conflict and by predicting the CRC scores using a combination of page metrics, the classifier is able to account for approximately 90% of the variation in scores. It is reasonable to assume that the topics covered on those pages with a high CRC are controversial and, therefore, topics for which more complex argument may occur.

The scope of controversy detection is broadened slightly in Choi et al. (2010) and (Awadallah et al., 2012) which both look at identifying controversy in news articles. In Choi et al. (2010), a controversial issue is defined as "a concept that invokes conflicting sentiments or views" and a subtopic as "a reason or factor that gives a particular sentiment or view to the issue". A method is proposed for the detection of controversial issues, based on the magnitude of sentiment information and the difference between the magnitudes for two different polarities. Firstly, noun and verb phrases are identified as candidate issues using a mixture of sentiment models and topical information.

The degree of controversy for these issues is calculated by measuring the volume of both positive and negative sentiment and the difference between them. For subtopic extraction, noun phrases are identified as candidates and, for these phrases, three statistical features (contextual similarity between the issue and a subtopic candidate, relatedness of a subtopic to sentiment, and the textual proximity of the issue and the candidate phrases) as well as two positional features, are calculated. The results for subtopic identification are poor, with an F-score of 0.50, however identifying controversial issues is considerably more successful, with a precision of 0.83[4].

Awadallah et al. (2012) present the OpinioNetIt system, which aims to automatically derive a map of the opinions-people network from news and other Web documents. The network is constructed in four stages. Firstly, generic terms are used to identify sample controversial topics. Next, opinion holders are identified for each topic, and their opinions extracted. The acquired topics and opinion holders are then used to construct a lexicon of phrases indicating support or opposition. Finally, this process is performed iteratively using the richer lexicon to identify more opinion holders, opinions and topics. Using this approach a precision of 0.72 is achieved in classifying controversial opinions.

Despite the specific domain limitations of this controversy detection work, (Dori-Hacohen and Allan, 2013) extends its scope to detecting controversy on the web as a whole, enabling users to be informed of controversial issues and alerted when alternative viewpoints are available. This is achieved by first mapping a given webpage to a set of neighbouring Wikipedia articles labelled on a controversiality metric, then combining the labels to give an estimate of the page's controversiality which is finally converted into a binary value using a threshold. This approach gives a 22% increase in accuracy over a sentiment-based approach, indicating that, although closely related, detecting controversy is more complex than simply detecting opinions and looking at where they differ.

Such widespread use of controversy detection offers the ability to address

[4]The precision is calculated based upon a user study where the participants are asked to confirm if an issue is controversial, as such, recall is not reported.

potential hotspot issues as they arise and the possibility of dealing with conflict in a debate at an early stage, before the quality of discussion can be be negatively impacted. Rumshisky et al. (2017), for example, take advantage of both content- and graph-based features to analyze the dynamics of social or political conflict as it develops over time, using a combination of measures of conflict intensity derived from social media data. Such methods for determining controversial issues can play a significant role in determining the argumentative structure inherent in a piece of text. Those points which are controversial are likely to attract not only more attention, but also a more even mix of supporting and attacking views, than those on which there is broad consensus. Lawrence et al. (2017b) make this connection explicit, showing how the divisiveness, or controversiality, of a proposition might be based upon the relative number of its supports and conflicts. A proposition with many of both might be taken to be divisive, whereas few of either might suggest only limited divisiveness. Alternatively, given a pair of propositions which are in conflict, the divisiveness of this conflict is shown to be a measure of the amount of support on both sides. It is easy to see how this process could be reversed, meaning that if we are able to identify controversial points in a piece of text, we already know something about the argumentative structure.

2.1.3 Citation Mining

Citation mining involves the labelling of citation instances in scientific writing with their rhetorical roles in the discourse. The techniques used to automatically determine the motivating factors behind each citation map closely to applications in argument mining, where text spans are labelled based on their argumentative role. For example, if a citation is being used to highlight a gap or deficiency in the referenced work, then the language used will be suggestive of conflict relations between the two; if a citation is being used to back up the current work, then there are likely argumentative support relations between the two.

There are a broad range of manual schemes for classifying citation motivation and citation function (the reason why an author chooses to cite a paper),

and (Teufel et al., 2006) looks at how this classification can be automated. A classification scheme is first developed using guidelines for twelve different categories (explicit statement of weakness, four types of contrast/comparison, six types of agreement/usage and neutral). Human annotators testing this scheme achieve a κ[5] of 0.72 and when implemented as an automatic procedure with the features listed below:

- Cue phrases
- Cues identified by annotators - 892 cue phrases identified by annotators (around 75 per category)
- Verb tense and voice used for recognising statements of previous/future/current work
- Location in paper/sentence/paragraph
- Self citations identified by author name

An accuracy of 0.77 and κ, determined based on level of agreement between the automated results and human annotated data, of 0.57 is achieved for classification to the 12 specified categories (or accuracy 0.83, κ 0.58 for 3-way classification positive/negative/neutral) based on an evaluation corpus of 116 articles, containing 2829 citations. Kappa is even higher for the top level distinction, collapsing the similar categories into just four (statement of weakness, contrast/comparison, agreement/usage and neutral) gives a κ value of 0.59. By comparison, the human agreement for this configuration is $\kappa = 0.76$. Whilst this leaves a significant gap between automated and human performance, it nevertheless suggests 'moderate agreement' using the automated approach, an encouraging result for a complex task.

An attempt to classify the opinion an author holds towards a work which they cite (for example, positive/negative attitudes or approval/disapproval) is

[5] κ is a statistical measure of inter-rater agreement, measuring pairwise agreement among a set of coders and correcting for expected chance agreement (Carletta, 1996). An interpretation of kappa values is offered by (Landis and Koch, 1977) which describes values between 0.01–0.20 as showing slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and 0.81–1.00 almost perfect agreement.

presented in Piao et al. (2007), where semantic lexical resources and NLP tools are used to create a network of opinion polarity relations. Sentences containing citations are extracted first, before determining the opinion orientation of the subjective words in the context of the citation. From these opinion orientations, the attitude of the author towards the work which they are citing is labelled.

Athar (2011) takes a similar approach, whereby analysis is performed on a corpus of scientific texts taken from the ACL Anthology, and consisting of 8,736 citations from 310 research papers manually annotated for their sentiment. Sentences are labelled as positive, negative or objective, with 1,472 used for development and training. Each citation is represented as a feature set in a Support Vector Machine (SVM) and processed using WEKA (Holmes et al., 1994) and the WEKA LibSVM library with the following features:

- **Word Level Features** Unigrams and bigrams as well as 3-grams to capture longer technical terms. POS tags are also included using two approaches: attaching the tag to the word by a delimiter, and appending all tags at the end of the sentence. A science-specific sentiment lexicon is also added consisting of 83 polar phrases such as efficient, popular, successful, state-of-the-art and effective.

- **Contextual Polarity Features** Sentence-based features e.g., presence of subjectivity clues which have been compiled from several sources along with the number of adjectives, adverbs, pronouns, modals and cardinals.

- **Dependency Structures** Typed dependency structures (De Marneffe and Manning, 2008) describing the grammatical relationships between words. For instance, in the sentence "CITE showed that the results for French-English were competitive to state-of-the-art alignment systems.", the relationship between results and competitive will be missed by trigrams but the dependency representation captures it in a single feature `nsubj_competitive_results`.

- **Sentence Splitting** each sentence is split by trimming its parse tree. Walking from the citation node towards the root, the subtree rooted at

the first sentence node is selected and the rest ignored

- **Negation** All words inside a k-word window of any negation term are suffixed with a token `_neg` to distinguish them from their non-polar versions.

The results show that 3-grams and dependencies perform best in this task with macro F-score 0.76 and micro F-score 0.89.

2.1.4 Argumentative Zoning

Argumentative Zoning (AZ) is the classification of sentences by their rhetorical and argumentative role within a scientific paper. For example, criticism or support for previous work, comparison of methods and results or goals. Although this approach of labelling a sentence by its role is slightly removed from the goal of identifying the argumentation structure contained within the document, it is clear that the information obtained by AZ provides a useful step towards determining the structure.

In Teufel et al. (2009), an annotation scheme covering fourteen possible roles is used to classify sentences into mutually exclusive categories. These categories extend the original seven categories presented in Teufel et al. (1999) and are designed to be applied to material from the life sciences domain as well as to the Computational Linguistics (CL) material considered in the earlier work. This categorisation highlights the link between Argumentative Zoning and Argument Mining. The 'AIM' (statement of specific research goal, or hypothesis of current paper) and 'OWN CONC' (findings, conclusions (non-measurable) of own work) categories, for example, are suggestive of conclusions. 'NOV ADV' (novelty or advantage of own approach) and 'SUPPORT' (other work supports current work or is supported by current work) suggestive of support relations, and 'GAP WEAK' (lack of solution in field, problem with other solutions) and 'ANTISUPP' (clash with somebody else's results or theory) suggestive of conflict relations.

Teufel et al. use a domain expert to encode basic knowledge about the subject, such as terminology and domain specific rules for individual cate-

gories, as part of the annotation guidelines. The produced guidelines include a decision tree, descriptions of the semantic nature of each category, rules for pairwise distinction of the categories and a large range of examples taken from both chemistry and computational linguistics. Human coders with background knowledge in computational linguistics, and varied experience in chemistry applied these guidelines, achieving inter-annotator agreement for chemistry with $\kappa = 0.71$ (N=3745, n=15, k=3). For CL, the inter-annotator agreement was $\kappa = 0.65$ (N=1629, n=15, k=3). As a comparison, the inter-annotator agreement for Teufel's original, CL-specific AZ with seven categories (Teufel et al., 1999) was $\kappa = 0.71$ (N=3420, n=7, k=3). This level of agreement between the three annotators is acceptable overall and supports the hypothesis that the task definition is domain-knowledge free. However, agreements involving the semi-expert are higher than the agreement between expert and non-expert, indicating that a general understanding of basic chemistry was not sufficiently adequate to ensure that the non-expert understood enough of the material to achieve the highest-possible agreement.

Merity et al. (2009) presents a maximum entropy classifier with each sentence of an article classified into one of the seven basic rhetorical structures from (Teufel et al., 1999). A maximum entropy model combined with the addition of new features to those used by Teufel gives an increase from 0.76 to 0.97 F-score on Teufel's Computational Linguistics conference paper corpus (48 computational linguistics papers, taken from the proceedings of the COLING, ANLP and ACL conferences between April 1994 and April 1995). The features used are described below:

- **Unigrams, bigrams and n-grams** Unigram and bigram features were included and reported individually and together (as n-grams). These features include all of the unigrams and bigrams above the feature cutoff.
- **First** The first four words of a sentence, added individually.
- **Section** A section counter which increments on each heading to measure the distance into the document.

- **Location** The position of a sentence between two headings (representing a section).
- **Paragraph** The position of the sentence within a paragraph.
- **Length** Length of sentence grouped into multiples of 3.
- **Teufel's (1999) features** To compare with previous work, most of the features that gave Teufel the best performance are also implemented.
- **Feature Cutoff** Instead of including every possible feature, a cutoff was used to remove features that occur less than four times.
- **History features** History features were used and Argumentative Zoning treated as a sequence labelling task with history lengths ranging from previous label to the previous four labels.

The results show that n-grams have by far the largest impact with a 21.39% reduction in accuracy when they are removed (the next largest impact being 1.24% for the first four words of the sentence). The history features also have an impact of just over 1%. It is shown that none of Teufel's individual features alone make a substantial contribution to the results when using the maximum entropy model. To evaluate the wider applicability of Argumentative Zoning, a corpus of Astronomy journal articles was also annotated with a modified zone and content scheme, and a similar level of performance (around 0.96 accuracy) was achieved.

2.2 Manual Argument Analysis

In this section we look at the task of manual argument analysis, considering the steps involved and tools available, as well as the limitations of manually analysing large volumes of text. Understanding manual analysis can offer unique insight into how this task can be automated and provides a valuable insight into how an analyst unpicks the complex argumentative relationships represented in natural language texts.

Although the argumentative structure contained within a piece of text (van Eemeren et al., 2014) can be diagrammed manually using pen and paper or simple graphics software, a wide range of specific argument diagramming tools (Scheuer et al., 2010) has been developed to allow an analyst to identify the argumentative sections of the text and diagram the structure which they represent (Kirschner et al., 2003; Okada et al., 2008). The advantages of this approach, as opposed to the use of non-specialised software, are discussed in Harrell (2005), though there is varied (and conflicting evidence of) impact on the the day-to-day activity within domains in which these tools are applied such as law, pedagogy, scientific writing (Lauscher et al., 2018a,b) and design (Scheuer et al., 2010). The majority of these tools, such as Araucaria (Reed and Rowe, 2004), Rationale (van Gelder, 2007), OVA (Bex et al., 2013) and Carneades (Gordon et al., 2007), require the analyst to manually identify the propositions involved in the argument being made and then connect them identifying the premises and conclusion. In many cases, this simple structure can then be extended with more specialised information depending on the nature of the analysis task being performed, for example, giving details of the Argumentation Schemes (Walton et al., 2008) used or details of the participants and their dialogical moves (for example, *questioning* or *asserting*) when analysing dialogue.

Generally, manual argument analysis, as carried out using the tools previously mentioned, can be split into four distinct stages as shown in Figure 2.1.

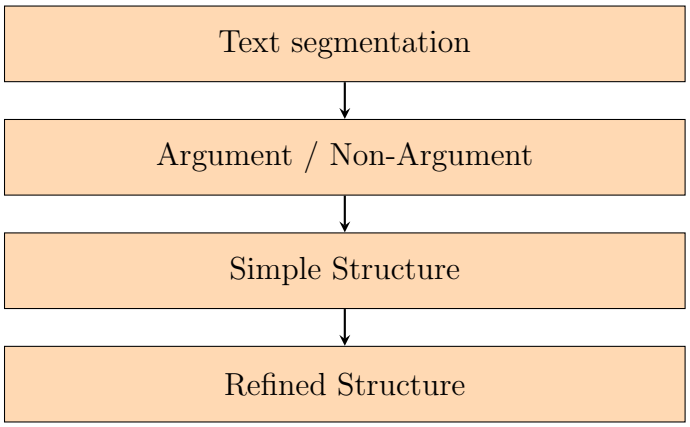


Figure 2.1: Steps in argument analysis

Though both manual and automated analysis techniques may develop a

more complex, hybrid approach in practice, the pipeline model presented here offers a good starting point from which to introduce the range of techniques currently available. Then in Section 2.3 we further dissect these steps, presenting a more detailed view of the individual argument mining steps and how they relate to the manual annotation process, explaining how increasingly the pipeline view oversimplifies complex interdependencies.

2.2.1 Text segmentation

Text segmentation involves the extraction of the fragments of text from the original piece that will form the constituent parts of the resulting argument structure. Text segmentation can be considered as the identification of a form of Elementary Discourse Units (EDUs). Though there are competing hypotheses about what constitutes an EDU (for example, (Grimes, 1975; Givón, 1983) view them as clauses; while (Hirschberg and Litman, 1993) views them as prosodic units; (Sacks et al., 1974) as turns of talk; (Polanyi, 1988) as sentences; and (Grosz and Sidner, 1986) as intentionally defined discourse segments), all agree that EDUs are non-overlapping spans of text corresponding to the atomic units of discourse. (Peldszus and Stede, 2013a) refers to these argument segments as 'Argumentative Discourse Units' (ADUs), and defines an ADU as a 'minimal unit of analysis', pointing out that an ADU may not always be as small as an EDU, for example, "when two EDUs are joined by some coherence relation that is irrelevant for argumentation, the resulting complex might be the better ADU" (p20).

Generally speaking, in argument analysis, the sections that the analyst extracts correspond to the propositions expressed explicitly by the text; however, some knowledge of the argument being made is often required in order to determine the exact boundaries of these propositions and how fine-grained the segmentation needs to be. In some cases, for example, propositional content can occur nested in reported speech, such as the sentence "Simon said this is a blue pen". The rest of the argument structure may refer to either the whole sentence ("Simon didn't say that"), to the statement "this is a blue pen" ("it's clearly a black pen") or to both parts separately ("Yes, I heard Simon say that,

but he's wrong, it's a black pen"). Another challenging example is *dislocation* which, similar to cleft constructions in syntax (Lasnik and Uriagereka, 1988), occurs when one segment is embedded into another, such as the example given in Saint-Dizier (2012): "Products X and Y because of their toxicity are not allowed in this building". In this case the conclusion, "Products X and Y are not allowed in this building", is split around the premise "because of their toxicity". As these examples show, robustly identifying the text segments required for an analysis can be challenging even for a human analyst.

An additional complication can occur in cases where some reconstruction of the argument is required in order to identify the points being made. There is a tendency for arguers to leave implicit an assumption required in order for their conclusion to follow from their premises. This can often occur when the omitted proposition is believed to be obvious; however it can also happen for a range of other reasons, for example, to increase the rhetorical force of the argument, or to conceal its unsoundness. Such missing premises are referred to as *enthymemes* (Hitchcock, 1985), and can cause difficulties for both automatic and manual segmentation due to the requirement of knowledge that may be outside the scope of that expressed in the text.

2.2.2 Argument / Non-Argument Classification

This step involves determining which of the segments previously identified are part of the argument being presented and which are not. For most manual analysis tools this step is performed as an integral part of segmentation: the analyst simply avoids segmenting any parts of the text that are not relevant to the argument. However, in some cases, for example where segmentation has been performed automatically or by a different analyst, this step must be carried out independently. In these cases the judgement as to whether a particular segment is argumentative can be made as a preliminary step in determining the structure, or left until the end of the analysis, when any segments left unconnected to the rest of the structure can simply be discarded.

Looking at the text shown in Example 1 below, we can see that the majority of Michael Buerk's introduction of Nick Dearden is non-argumentative, with

only the single claim identified that Mr Dearden would like people not to have to pay their debts. Meanwhile, almost the entirety of the response (excluding brief connectives) forms part of the argument structure.

> **Michael Buerk:** John Lamiday, thank you very much indeed for joining us this evening. Our third witness is Nick Dearden, who is director of the Jubilee Debt Campaign. *Mr Dearden, you'd like people not to have to pay their debts*. Where's the morality in that?
>
> **Nick Dearden:** *I wouldn't like people not to have to pay their debts across the board*. But I think what we say is that *this isn't simply a matter of individual morality*. *Debt is used time and again as a set of economic decisions, and political decisions, to achieve certain things in society*. And *very often what high levels of debt can mean, and especially when the debt is on very unjust terms, is a massive redistribution of wealth in society, from the poorest to the richest*.

Example 1: Excerpt from the BBC Moral Maze 'Money' corpus (`http://corpora.aifdb.org/Money`). Argumentative segments are *highlighted*.

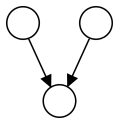
In some cases however, this task can be remarkably demanding. 'Letters to the Editor' contributions, for example, can sometimes offer rich pickings for the argument analyst, but such letters can often be little more than frivolity or wit masquerading as argument and inference. Distinguishing argument from non-argument in this domain is extremely demanding, even for a highly trained human analyst.

2.2.3 Simple Structure

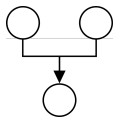
Once the elements of the argument have been determined, the next step is to examine the links between them. This can be as simple as noting segments that are thematically related, but usually involves the identification of support and attack relations between segments. Whilst these relations can be simply

labelled pairs, it is common to consider the varying ways in which components can work together (Groarke et al., 1997):

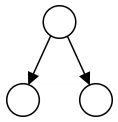
Convergent Arguments In a convergent argument, multiple premises are used to to independently support a single conclusion. In this case the premises act on their own and the removal of one premise from the argument does not weaken the others. From Example 1 we can see that "what we say is that this isn't simply a matter of individual morality" and "I wouldn't like people not to have to pay their debts across the board" independently support "Mr Dearden would like people not to have to pay their debts".



Linked Arguments In a linked argument, multiple premises work together to support a conclusion. The important point here is that each premise requires the others in order to work fully. In Example 1, the statements "Debt is used time and again as a set of economic decisions, and political decisions, to achieve certain things in society" and "very often what high levels of debt can mean, and especially when the debt is on very unjust terms, is a massive redistribution of wealth in society, from the poorest to the richest" work together to support the point "what we say is that this isn't simply a matter of individual morality".



Divergent Arguments In some cases the same premise may support multiple conclusions. Divergent arguments are somewhat less common and, as such, are not supported by those analysis tools which, for example, are limited to analysing arguments in a tree structure.



Though Example 1 does not include a divergent argument, Dearden might have said, 'And if it's not individual morality, then the state should take some of the responsibility,' which would have offered a second conclusion based on the premise of individual morality.

Sequential (or Serial) Arguments The final way in which multiple premises can support a conclusion is in a sequential argument. In this case, one premise leads to another and this, in turn, leads to the conclusion. In Example 1, the statements "very often what high levels of debt can mean, and especially when the debt is on very unjust terms, is a massive redistribution of wealth in society, from the poorest to the richest", "what we say is that this isn't simply a matter of individual morality" and "Mr Dearden would like people not to have to pay their debts" follow a sequential structure.

Hybrid Argument Structure More complicated arguments, such as that in Example 1, usually involve several instances and combinations of the above elements into a larger, hybrid, argument structure. The complete analysed structure of Example 1 can be seen in Figure 2.2.

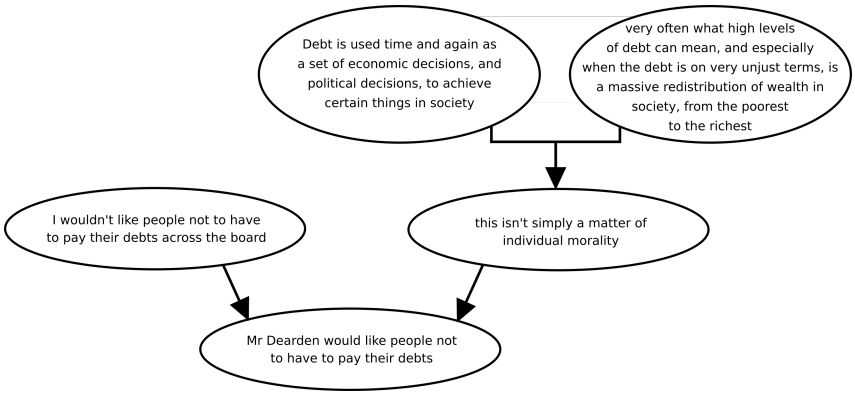


Figure 2.2: Simple argument structure of the text in Example 1

We must also consider conflict, or attack, relations between propositions. These include both standard conflict relations where one proposition directly conflicts with another, as well as more complex forms of defeating an argument (Pollock, 1986):

Rebutting Attacks Rebutting arguments express a position that is directly incompatible with a conclusion (Pollock, 1986, p.38). Later in the debate from which Example 1 is drawn, an opponent, Michael Portillo, says, 'People who lend money, that is to say, people who save money, say through building societies, are very ordinary people.' This offers a direct, rebutting attack to

Dearden's conclusion, expressed in Example 1, that debt is a massive redistribution of wealth from the poorest to the richest.

Undercutting Attacks Undercutting arguments attack or conflict with the inference between a premise and a conclusion, and, as such, offer a reason for no longer believing the conclusion, rather than for believing the negation of the conclusion (Pollock, 1986, p.39). Though the fragment of debate from which Example 1 is drawn does not offer clear examples of undercutting, Portillo might have retorted with, 'If there were politicial decisions being taken, they are being taken by elected officers – so state actions don't require more than individual morality'. Such an attack does not directly counter the conclusion, but instead focuses on the robustness of the passage from premise to conclusion.

Although this approach to identifying argument structure is by far the most common, other methodologies, such as (Toulmin, 1958) are also widely used; perhaps the clearest synthesis for computational purposes is presented by the philosopher J.B. Freeman (Freeman, 1991, 2011). For argument mining, successful extraction of argument structure in one form can often be translated into others, modulo expressivity constraints (we discuss different argument representations and formats as well as the translation between them in Chapter 3.

2.2.4 Refined Structure

Having determined the basic argumentative structure, some analysis tools allow this to be refined further. For example, Araucaria, Carneades, Rationale and OVA allow the analyst to identify the argumentation scheme related to a particular structure. Argumentation schemes are patterns of inference, connecting a set of premises to a conclusion, that represent stereotypical patterns of human reasoning. Such schemes were originally viewed as rhetorical methods by which a speaker could influence their audience; later they have also been adopted as a way to distinguish good arguments from bad. Argumentation schemes can thus be seen as a historical descendant of the topics of

Aristotle (1958), and, much like Aristotle's topics, play a valuable role in both the construction and evaluation of arguments. Arguments are evaluated based on a set of critical questions corresponding to the scheme which, if not answered adequately, result in the argument to which the scheme corresponds defaulting.

The 'Argument from Expert Opinion' scheme (Walton, 1996) is commonly used to illustrate the concept:

Major Premise: Source E is an expert in subject domain S containing proposition A.

Minor Premise: E asserts that proposition A is true (false).

Conclusion: A is true (false).

with the associated critical questions:

1. *Expertise Question:* How credible is E as an expert source?
2. *Field Question:* Is E an expert in the field F that A is in?
3. *Opinion Question:* What did E assert that implies A?
4. *Trustworthiness Question:* Is E personally reliable as a source?
5. *Consistency Question:* Is A consistent with what other experts assert?
6. *Backup Evidence Question:* Is E's assertion based on evidence?

Recent study has resulted in the identification and analysis of the most important and commonly used schematic structures (Hastings, 1963; Perelman and Olbrechts-Tyteca, 1969; Kienpointner, 1992; Pollock, 1995; Walton, 1996; Grennan, 1997; Katzav and Reed, 2004; Walton et al., 2008). Whilst there is much overlap in these classifications, they often differ in their granularity: Pollock identifies fewer than ten schemes; Walton, nearly thirty; Grennan, more than fifty; and Katvaz & Reed, more than one hundred. Due to these differences, it is common for analysis tools to retain the grouping of schemes into sets. Araucaria, for example, supports the Walton, Grennan, Perelman & Olbrechts-Tyteca, Katzav & Reed and Pollock scheme sets.

Experiments on the annotation of Walton schemes by annotators with a strong background in linguistics but who were provided with only the description of the schemes given in Walton et al. (2008) have shown that this is an exceptionally difficult task, with results differing in both numbers of arguments annotated and the distributions of units (Lindahl et al., 2019). However, recent developments in annotation guidelines for these schemes, including the decision tree based method described in Lawrence et al. (2019a), suggest that this situation can be improved and offer hope for the construction of scheme annotated corpora.

2.2.5 Limitations of manual analysis

Although these tools can be used for the analysis of small sections of text, analysing large volumes of text quickly and, certainly in anything approaching real time, is beyond their scope. Compendium[6] IBIS map facilitators are the closest, but the analysis involved is at a much higher level. The major limitation is the amount of information that can be handled by a single analyst. Efforts have been made to overcome this obstacle by both crowdsourcing of annotation (Ghosh et al., 2014) and using hardware designed to allow multiple trained annotators to collaborate on the same analysis (Bex et al., 2013). In the first case, by applying a clustering technique to identify which pieces of text were easier or harder for trained experts to annotate, it was determined that the crowdsourced results were only accurate for those segments that were identified as being easier for expert annotators. In the second case, whilst the AnalysisWall, a touchscreen measuring 11 feet by 7 feet running bespoke analysis software (Bex et al., 2013), has been used to analyse several hour-long radio programmes in real time, it still does not come close to allowing for the analysis of the vast volumes of data produced every day.

[6] `http://compendium.open.ac.uk/`

2.3 Argument Mining: Automating Argument Analysis

In the preceding sections, we have looked first at a range of different techniques which are precursors to the task of argument mining, and at the manual analysis of the argumentative structure of a text, gaining an understanding of both the nature of argumentative structure as well as the process by which a human analyst understands and extracts this structure. In this section we now break down the argument mining task into a range of individual challenges (see Figure 2.3). In Sections 2.4, 2.5 and 2.6, we will then look at each of these tasks in more detail, drawing together work targeted at varying domains, and using different approaches, to understand the challenges and progress made in each of these areas.

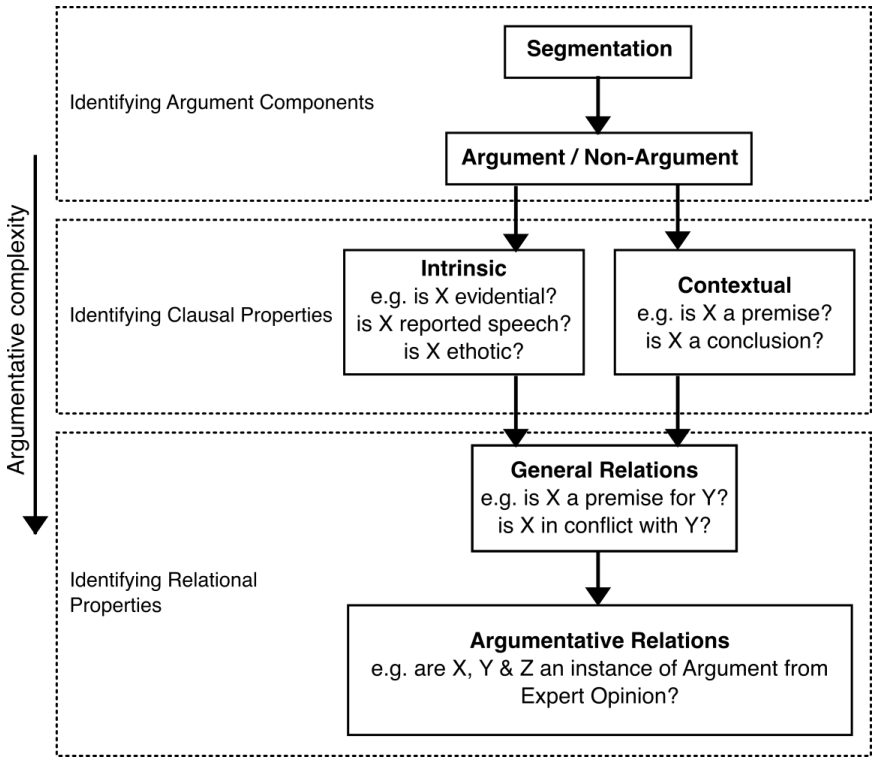


Figure 2.3: The tasks and levels of complexity in argument mining techniques

For the purposes of this review, we use these tasks as a framework to present and organise the work carried out in the field. In Section 2.4 we look at automatic approaches for identifying argument components and determining their boundaries. In Section 2.5 we move on to look at the automatic

identification of properties which these clauses have and in Section 2.6 we look at the identification of relations from simple premise/conclusion relations to argumentation scheme instances and dialogical properties. Where a piece of work offers a large contribution to several areas, we include these in multiple sections, grouping each part of their contribution with other works addressing the same tasks individually. For each task, we consider work carried out using a broad range of techniques, including statistical and linguistic methods.

We have seen in Section 2.2 how the steps in manual analysis increase in complexity from segmenting argumentative components to identifying argumentation schemes and dialogical relations. These levels are also reflected in the automation of argument analysis. In some cases it is sufficient to know merely the range of argumentative types used in order to grade student essays (Ong et al., 2014), to know what stance an essay takes towards a proposition in order to check it provides appropriate evidence to back-up its stance (Persing and Ng, 2015), or whether a claim is verifiable in order to flag these in online discussions (Park and Cardie, 2014). However, if the goal is to reconstruct enthymemes (Razuvayevskaya and Teufel, 2017) (see also the discussion of (Feng and Hirst, 2011) in Section 2.6.2) or ask critical questions about support relations, we also need to extract the nature of the argumentation schemes being used.

In Figure 2.3, we show how these automatic tasks are inter-related. Starting from the identification of argument components by segmenting and classifying these as part of the argument being made or not (these tasks are sometimes performed simultaneously, sometimes separated and sometimes the latter is omitted completely), we move down through levels of increasing complexity: first considering the role of individual clauses (both intrinsic, such as whether the clause is reported speech, and contextual such as whether the clause is the conclusion to an argument); secondly considering argumentative relations from simple premise/conclusion relationships,; and thirdly whether a set of clauses forms a complex argumentative relation, such as an instance of an argumentation scheme. A similar classification of argument mining tasks is given in Cabrio and Villata (2018), with 'Component Detection' being split into

the subtasks of 'Boundary Detection' and 'Sentence Classification'. Whilst this represents a robust starting point, it is also important to distinguish the types of classification (argument/non-argument and intrinsic/contextual). Cabrio and Villata also include the broad categorisation of 'Relation Prediction', which again can be further broken down, looking at both general and argumentative relations.

The arrows shown between tasks in the figure indicate ways in which the results from one task have been used to inform the execution of another. For example, the arrow from the "Argument/Non-Argument" task to the "Contextual Clausal Properties" task, reflects much early argument mining work (e.g. (Moens et al., 2007)) which performed these tasks in sequence; deciding which parts of the text were argumentative and then assigning a role to them. This approach has been challenged however, with Carstens and Toni (2015) being the first to point out that whether a sentence is argumentative or not often depends on the context in which it is used, and instead advocating classifying relations first and then considering sentences to be argumentative if they have a relation connecting them (reflected in the arrow from "General Relations" to "Argument/Non-Argument").

Similarly, some tasks can inform each other, for example, where Feng and Hirst (2011) showed that argument scheme instances could be classified given general relations between ADUs, Lawrence and Reed (2015) showed that such general relations can be determined by classifying argument scheme components directly from segmented text. This inter-dependency between tasks has given rise to a growth in the application of multi-objective learning approaches (e.g. (Eger et al., 2017; Hou and Jochim, 2017; Galassi et al., 2018; Morio and Fujita, 2018)), where all tasks are learnt and performed at the same time. These examples highlight how the simple pipeline view of argument mining, which characterises a lot of older research work, is increasingly being superceded by more sophisticated and interconnected techniques.

Developments in argument mining are both being informed by, and informing, the related areas discussed in Section 2.1. For example, the work of Ong et al. (2014) closely parallels both argumentative zoning and citation mining,

offering the opportunity to link related elements automatically identified in scientific writing, such as how a claim may be supported by a nearby citation. Rumshisky et al. (2017), look at the dynamics of social or political conflict as it develops over time, automatically identifying controversial issues where such conflict is occurring. While, Accuosto and Saggion (2019) show how argumentation in certain sections of a publication (in this case abstracts), can be a good indicators of the quality of the work as a whole.

2.4 Identifying Argument Components

The automatic identification of the argumentative sections of a text corresponds to the process of argument/non-argument classification discussed in Section 2.2.2. Whilst carrying out this task in isolation does not give us a detailed picture of the argument structure, it has found use in, for example, predicting the usefulness of online reviews based solely on the *amount* of argumentative text which they contain Passon et al. (2018).

One of the first approaches to argument mining, and perhaps still the most pioneering, is the work carried out by Moens et al. (Moens et al., 2007; Palau and Moens, 2009; Mochales and Moens, 2011), which first attempts to detect the argumentative parts of a text by first splitting the text into sentences and then using features of these sentences to classify each as either "Argument" or "Non-Argument". By training a range of classifiers on manually annotated examples from the Araucaria corpus (Reed, 2006), an accuracy of 0.74 is obtained using a multinomial naïve Bayes classifier trained on word couples, verbs and text statistics.

Similarly, (Goudas et al., 2014) looks at extracting arguments from social media proposing a two-step approach for argument extraction similar to that used by Moens et al., first employing a statistical approach through the use of machine learning and more specifically, the logistic regression classifier, to classify sentences as being part of the argument being made or not. This approach is applied to a corpus obtained from social media, concerning renewable energy sources in the Greek language, and for identifying sentences that contain

arguments, an increase in performance from an F-score of 0.21, for the base case, to 0.77 is achieved. This approach is further developed in Sardianos et al. (2015), where Conditional Random Fields are used to identify those segments from similar Greek social web texts which contain argumentative elements.

Although these results are encouraging, it is worth noting that the classification of sentences carried out refers only to features intrinsic to the sentence and as such the classification is not robust for sentences which may be part of an argument in one context, but not in a different context. Several examples of sentences that can be viewed as argumentative in some contexts, but not in others, can be seen in Carstens and Toni (2015), which instead advocates classifying pairs of sentences according to their argumentative relation and, if the relation is classified as support or attack, considering both sentences to be argumentative. In Section 2.6 we look at such techniques for identifying relations, and show that Carstens and Toni's approach is in many cases preferable to the pre-identification of argumentative components.

Saint-Dizier (2018) offers an example of a situation where domain knowledge is required in order to determine whether or not a proposition is argumentative. Given the issue "Vaccine against Ebola is necessary" it is argued that the proposition "7 people died during Ebola vaccine tests" is irrelevant or neutral with respect to the issue under a knowledge-based analysis, whereas a naïve reading would rather interpret it as an attack. The importance of contextual domain knowledge highlighted by this example was first explored by Saint-Dizier in Saint-Dizier (2017) where, via the analysis of various corpora, the types of knowledge that are required to develop an efficient argument mining system, are explored. This exploration shows that, in about 75% of cases, some contextual knowledge is required to accurately identify arguments with respect to a controversial issue.

The idea that the context in which a text span appears can determine whether it is part of an argument or not (Opitz and Frank (2019) have shown that context can be more important than content), can be problematic for the general application of the supervised machine learning approaches discussed so far. In cases where context is not adequately captured, a model trained

on one set of data can struggle to classify spans in another set of data where the context is different. As a result, rule-based and unsupervised learning approaches have also been applied to this task. The application of an unsupervised extractive summarisation algorithm, TextRank, for the identification of argumentative components is explored in Petasis and Karkaletsis (2016). The motivation is to examine whether there is any potential overlap between extractive summarisation and argument mining, and whether approaches used in summarisation (which typically model a document as a whole) can have a positive effect on tasks of argument mining. Evaluation is performed on two corpora containing user posts from an on-line debating forum and persuasive essays, with results suggesting that graph-based approaches and approaches targeting extractive summarisation can have a positive effect on tasks related to argument mining.

Similarly, Wachsmuth et al. (2017b) propose a model for determining the relevance of arguments using PageRank (Brin and Page, 1998). In this approach, the relevance of an argument's conclusion is decided by what other arguments reuse it as a premise. These results are compared to an argument relevance benchmark dataset, manually annotated by seven experts. On this dataset, the PageRank scores are found to beat several intuitive baselines and correlate with human judgments of relevance.

One of the first supervised learning approaches to segmentation was introduced by (Soricut and Marcu, 2003) as a part of the `SPADE` system, which also operates on lexicalised syntactic trees. The authors compute the probability of inserting a discourse boundary between a child and parent node and attained an F-score of 0.83.

The current state-of-the-art results for EDU identification are obtained by the two-pass system of Feng and Hirst (2014), which uses a sequence labelling approach. Similar to Soricut and Marcu (2003), the method makes predictions over pairs of tokens that are enriched with syntactic features. Feng and Hirst showed that predicting over token pairs and making these predictions in two passes improves the results, achieving a 0.93 F-score on the recognition of in-sentence boundaries.

ADU identification however, is considerably more challenging than identifying EDUs, requiring an understanding of the argumentative function of each span. Madnani et al. (2012), aims to separate argumentative discourse into two categories; firstly, argumentative text, used to express claims and evidence, and secondly language used to present and organise the claims and evidence ("shell"). In the example sentence "So I think the lesson to be drawn is that we should never hesitate to use military force...to keep the American people safe", the underlined text is identified as shell. Separating shell from argumentative text is attempted using three methods: a rule-based system, a supervised probabilistic sequence model, and a principled hybrid version of the two. The rule-based system gives an F-score of 0.44, with the hybrid version giving 0.61 compared to 0.74 for a human annotator and 0.21 for a baseline which labels words as shell if they appear frequently in persuasive writing. The rule-based system uses a set of 25 hand-written regular expression patterns for example, "I [MODAL] [ADVERB] AGREEVERB with the AUTHORNOUN". The Supervised Sequence Model is based on conditional random fields (CRFs) using a small number of general features based on lexical frequencies with the intuition behind these features being that shell language generally consists of chunks of words that occur frequently in persuasive language. It is important to note that, although the material identified as shell is not a part of the argument being made, this material contains valuable information about the argument structure, often indicating the occurrence of certain speech acts, or containing discourse markers (Hutchinson, 2004).

Lawrence et al. (2014) present an alternative supervised learning approach to ADU segmentation, focusing specifically on identification of ADU boundaries. Two naïve Bayes classifiers are used to perform *Proposition Boundary Learning*, one to determine the first word of a proposition and one to determine the last. The classifiers are trained using a set of manually extracted propositions as training data. The text to be segmented is first split into words and a list of features is then determined for each of these words. The features used cover both intrinsic (the word itself, its length, and Part Of Speech) and contextual (the word/punctuation before and the word/punctuation after). By

looking at more general features (length and POS) and contextual features, this approach aims to overcome the variability in specific words that may start (or end) a proposition.

Having trained the classifiers, this same list of features is then determined for each word in the test data, enabling the classifiers to label each word as being 'start' or 'end'. Once the classification has taken place, the individual starts and ends are matched to determine propositions, using their calculated probabilities to resolve situations where a start is not followed by an end (i.e. where the length of the proposition text to be segmented is ambiguous). Using this method, a 32% increase in accuracy is achieved over simply segmenting the text into sentences when compared to argumentative spans identified by a manual analysis process.

Ajjour et al. (2017) also find that considering the broader context of surrounding words, or even the document as a whole aids in locating proposition boundaries. The approach in this case is framed as a sequence labelling task, with a neural network model utilising structural, syntactic, lexical and pragmatic features, as well as capturing long- distance dependencies. Capturing the entire text with this model provides the best results across all domains, with F-scores of up to 0.89.

Even reliably identifying ADU segment boundaries, however, is being recognised as insufficient for identifying ADUs simply because ADUs typically express propositions with a variety of linguistic surface phenomena obfuscating that propositional content. Mood, anaphora, ellipsis, deixis, reported speech and more all introduce new challenges for ADU identification. (Jo et al., 2019) have used a combination of techniques, some statistical, some rule-based and some hybrid, organised in a cascade structure, in order to attempt to recover the propositional structure underlying ADUs, in order to improve the performance of other argument mining tasks.

2.5 Automatic Identification of Clausal Properties

In the previous section we explored a range of techniques for identifying the sections of a text which are argumentative, however, this does not yet tell us anything about the nature of these argumentative text spans, or how they work together. We now move on to look at techniques for automatically identifying properties of argumentative components. In this section, we look at identifying the function of each text span, firstly considering intrinsic properties (e.g. whether a text span is evidential in nature) and then look at contextual clausal properties, describing how a text span is used in the argument as a whole (e.g. as a premise or conclusion). In Section 2.6, we move on to look at the identification of inter-clausal relations, for example, given a pair of text spans, identifying any support or conflict relationship between them.

2.5.1 Intrinsic Clausal Properties

The first type of properties we look at are those which are intrinsic to the text span itself. Whilst these properties are limited in what they tell us about the overall argumentative structure, they provide valuable information about the role that a particular text span is playing in the argument as a whole. For example, knowing that a text span constitutes a verifiable claim suggests a link to a piece of evidence in the text supporting this claim (Park and Cardie, 2014), knowing that a text span is increasing the author's ethos suggests that it is supporting a specific argument which they are making (Duthie et al., 2016a), knowing the type of evidence provided can be used to assign different weights to statements in clinical trials (Mayer et al., 2018), or help understand rulings in disability benefits claims (Walker et al., 2018).

Verifying the acceptability of text spans used as premises in an argument is a central issue in the linguistic and philosophical study of argumentation (Freeman, 2000). In the study of persuasive communication and rhetoric, this has led to a variety of typologies of evidence. For example, Reynolds and Reynolds (2002) distinguish between statistical, testimonial, anecdotal and

analogical evidence; while Hoeken and Hustinx (2003) use a revised distinction between individual examples, statistical information, causal explanations, and expert opinions; and Fahnestock and Secor (1988) employ the classical stasis issues of fact, definition, cause, value, and action.

This diversity is also evident in the computational classification of propositions and evidence. In Park and Cardie (2014), online user comments are examined for propositions that are UNVERIFIABLE, VERIFIABLE NON-EXPERIENTIAL, or VERIFIABLE EXPERIENTIAL with associated supports of type *reason*, *evidence*, and *optional evidence*, respectively. A proposition is considered verifiable if it contains an objective assertion with a truth value that can be proved or disproved with objective evidence. Verifiable propositions are further split into experiential or non-experiential depending on whether or not the proposition is about the writer's personal state. For example, "My son has hypoglycemia" is tagged as Verifiable Experiential, whereas "food allergies are seen in less than 20% of the population" is marked as Verifiable Non-Experiential. Following an annotation scheme developed on 100 randomly selected comments, manual annotation inter-coder reliability is moderate, yielding an Unweighted Cohen's κ of 0.73 whilst Support Vector Machine classifiers trained with a range of features including n-grams and features specific to each class, exhibit statistically significant improvement over the unigram baseline, achieving a macro F-score of 0.69. These results show that identifying propositions of these types can be achieved with reasonable accuracy, however this would still need to be developed in order to identify the relations between these propositions and determine the argument structure. By having an indication of the required support for each proposition, this structure could then be used to identify areas where a proposition is not adequately supported.

These classifications are revised in Park and Cardie (2018) to: propositions of non-experiential fact (*fact*); propositions of experiential fact (*testimony*); propositions of value (*value*); propositions of policy (*policy*); and reference to a resource (*reference*). With these revised proposition categories and their associated supports of type *reason* and *evidence*, a further annotation study was carried out, resulting in the Consumer Debt Collection Practices (CDCP)

corpus. This corpus consists of 731 user comments on the Consumer Debt Collection Practices ruling, with 4931 elementary units (of which the majority were propositions of value - 45%), and 1221 support relations (1,174 reason, and only 46 evidence). On this dataset, Niculae (2018) achieved a maximum F1-Score of 0.74 for proposition classification using linear structured SVMs.

Egawa et al. (2019) adjust the annotation scheme of Park and Cardie slightly, replacing *reference* with *rhetorical statement* (which implicitly states the subjective value judgement by expressing figurative phrases, emotions, or rhetorical questions) and replacing the relations with the more standard *attack* and *support*. This scheme was then used to annotate 345 posts from the ChangeMyView sub-reddit[7] resulting in 4,612 proposition classifications and 2,713 relations which were then used in analysing the semantic role of persuasive arguments.

The value of being able to identify verifiable propositions is highlighted by the classification of evidence types presented in Addawood and Bashir (2016), where Twitter posts are automatically identified as either a News media account (NEWS), Blog post (BLOG) or No Evidence (NO EVIDENCE). The data for this study is taken from tweets on the FBI and Apple encryption debate, with 3000 tweets annotated. Support Vector Machines (SVM) classifiers trained with n-grams and other features capture the different types of evidence used in social media and demonstrate significant improvement over the unigram baseline, achieving a macro-averaged F-score of 0.83. Similarly, Dusmanu et al. (2017) look at argumentative tweets classifying them as either fact or opinion with an F-score of 0.80 and the source of their information (e.g., CNN) with an F-score of 0.67.

The classification of factual statements for critical evaluation has gained prominence as part of fact-checking. Hassan et al. (2015) classify sentences as non-factual, unimportant factual, and check-worthy factual. Similarly, Patwari et al. (2017) and Jaradat et al. (2018) automatically determine the fact-check-worthiness of factual claims in political debates. Naderi and Hirst (2018a) automatically distinguish between true, false, stretch, and dodge statements

[7] `https://www.reddit.com/r/changemyview/`

in parliamentary proceedings.

Anand et al. (2011) consider a different level of intrinsic clausal properties than those discussed so far, looking not at properties related to their verifiability, but at their persuasive function. This work describes the development of a corpus of blog posts where attempts to persuade and the corresponding tactics employed in this persuasion are annotated. Persuasion involves the change in mental state of the other party classed as either 'Belief Revision', 'Attitude Change' or 'Compliance Gaining'. The methods that can be used to achieve these changes in mental state are considered in Marwell and Schmitt (1967), which offers twelve strategy types for securing behavioural compliance. A further six non-logical "principles of influence" are covered in Cialdini (2001). By combining these with argumentative patterns inspired by (Walton et al., 2008), and removing overlapping tactics, Anand et al. produce a list of 16 types of rhetorical tactic for persuasive acts. By using a naïve Bayes Classifier for seven possible combinations of three feature sets to perform this classification, Anand et al. report a best result with an F-score of 0.58. However, rhetorical relations are often implicit and not clearly indicated in the text, and as such, their discovery requires a richer set of features.

Duthie et al. (2016a) consider another facet of persuasion, using a pipeline of techniques to extract positive and negative ethotic statements (Aristotle, 1991) (those relating to the character of a person) from parliamentary records. Whilst this work differs from many other argument mining approaches (which despite often looking at persuasion, nonetheless typically focus exclusively on *logos* rather than *ethos* or *pathos*) there is a clear link, with ethotic relations often following the same logotic structures, but with the character of a person as their target. In this work, those statements in which the speaker refers to the character of another person (referred to as Ethotic Sentiment Expressions, *ESE*s) and those in which they do not (*non-ESE*s) are first extracted using a combination of Named Entity Recognition, Part-Of-Speech tagging and a set of domain specific rules to locate statements referring to another person, organisation or agentive entity. These are then passed to the anaphora layer where both source-person and target-person of the statement are retrieved

from the original text. Finally, a sentiment layer consisting of a sentiment classifier combined with sentiment and ethotic word lexicons classifies ESEs as positive and negative. The resulting pipeline achieves an F-score of 0.70 for ESE/non-ESE classification compared to 0.45 for a baseline classifier which predicts only the target class (ESE), and 0.78 for +/-ESE classification, compared to a baseline of 0.67. A similar corpus of statements aimed at defending against ethotic attacks, or defending the speaker's reputation, is presented in Naderi and Hirst (2018b), and extracted from various issues in Canadian parliamentary proceedings.

In Villalba and Saint-Dizier (2012), an approach to the identification and analysis of arguments as they appear in opinion texts is developed. Examples are given that show that arguments are either: incorporated into evaluative expressions with a heavy semantic load (for example, evaluative adjectives such as 'repas familial' means a meal that has properties such as casual, homemade, good and abundant), or, composed of an evaluation and one or more discourse structures such as justification, elaboration or illustration whose aim is to persuade the reader of the evaluation.

For example:

- **Justification**: *The hotel is 2 stars* [JUSTIFICATION *due to the lack of bar and restaurant facilities*].
- **Reformulation**: *Could be improved* [REFORMULATION *in other words, not so good*].
- **Elaboration by Illustration or Enumeration**: *The bathrooms were in a bad condition*: [ILLUSTRATION *the showers leaked, and the plug mechanism in the bath jammed...*] *Breakfast selection is very good* [ENUMERATION *with a range of cereals, tea and coffee, cold meats and cheese, fresh and canned fruit, bread, rolls and croissants, and a selection of cooked items.*]
- **Elaboration via Precision**: *Friendly and helpful staff*, [PRECISION *especially the service executives at the counter.*]

- **Elaboration via Comparison**: *These head phones are excellent,* [COMPARISON *as if you are in a concert room.*]
- **Elaboration via Consequence**: *a high soundproofing* [ELAB–CONSEQUENCE *that allows you to have a rest after a long working day*]
- **Contrast**: *The price is very reasonable* [CONTRAST *but comfort is rather poor.*]
- **Concession**: *Very quiet* [CONCESSION *in spite of its downtown location in a nightlife area.*]

These relations are processed using TextCoop (Saint-Dizier, 2012), a platform designed for discourse analysis, with a logic and linguistic perspective. The results compared to a manual annotation on a corpus of 50 texts range between precision = 0.85-0.92, recall = 0.76-0.86, over the eight relations listed above.

The Automatic Argumentative Analysis (A3) algorithm described in Pallotta and Delmonte (2011) provides an alternative approach to classifying statements according to rhetorical roles. A3 is a module developed based on the GETARUNS system (Delmonte, 2007) for Interaction Mining (the discovery and extraction of insightful information from digital conversations, namely those human-human information exchanges mediated by digital network technology). The module takes as input the complete semantic representation produced by GETARUNS and produces argumentative annotation using the following 20 discourse relation labels: circumstance, narration, adverse, obligation, evaluation, statement, result, hypothesis, elaboration, permission, cause, motivation, explanation, agreement, contrast, question, inception, setting, evidence and prohibition. These labels come partly from Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) and partly from other theories, including those reported by Hobbs (Hobbs, 1993) and Dahlgren (Dahlgren, 1988).

Discourse relations are automatically extracted by GETARUNS and these are then mapped onto five Meeting Description Schema (MDS), (Pallotta et al.,

2004) argumentative labels: ACCEPT, REJECT/DISAGREE, PROPOSE/-SUGGEST, EXPLAIN/JUSTIFY and REQUEST. In the training stage, the system was used to process the first ten dialogues of the International Computer Science Institute (ICSI) meetings corpus (Janin et al., 2003) containing a total number of 98,523 words and 13,803 turns. In the test stage, two different dialogues were randomly chosen to assess the performance of the A3 algorithm and on a total of 2,304 turns, 2,247 received an automatic argumentative classification, yielding a recall value of 0.98 (precision 0.81, F-Score 0.89).

Having labelled text segments in this way, it is easy to visualise them using, for example, conversation graphs (Ailomaa and Rajman, 2009). Conversation graphs are diagrams that summarise what topics were discussed, how long they were discussed, which participants were involved in the discussion and what type of arguments they contributed (an example conversation graph can be seen in Figure 2.4). Conversation graphs can be built directly by looking at the MDS labels assigned to a conversation's turns.

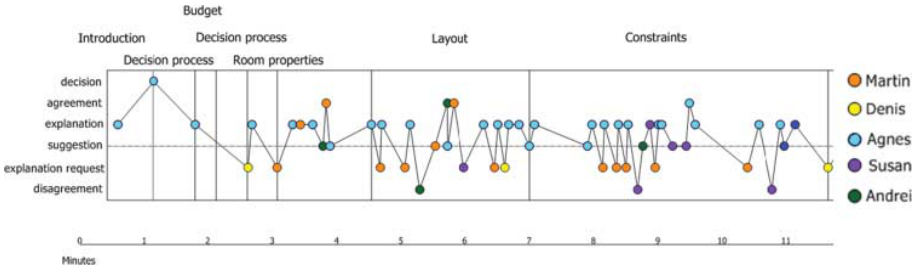


Figure 2.4: Conversation graph from (Ailomaa and Rajman, 2009)

The benefits of using even a simple linguistic analysis to study the argumentative structure of a document are illustrated in Ong et al. (2014) where a series of simple rules are used to tag sentences with their role (either Current Study, Hypothesis, Claim, or Citation), for example, if the sentence contains a four-digit number, then it is tagged as 'Citation', if the sentence contains string prefixes from {suggest, evidence, shows, essentially, indicate}, then it is tagged as 'Claim'. This approach again highlights the similarities between Argumentative Zoning (Section 2.1.4) and the determination of argumentative role. The ability to determine these roles offers the opportunity to link related

elements, for example a Claim may be backed by a nearby Citation.

Wyner et al. (2012) also use simple linguistic cues, in this case to support manual analysis by providing a rule-based tool for supporting textual analysis by semi-automatic identification of argumentative sections in the text. The tool is aimed specifically at online product reviews, and highlights potential argumentative text in the review according to discourse indicators (explicitly stated linguistic expressions of the relationship between statements (Webber et al., 2011)) and terminology specific to the domain (for example, product names and their properties). The tool uses a set of discourse indicators, sentiment terminology, a user model, and a domain model. Discourse indicators are used to locate premises (after, as, because, for, since, when, assuming,...), conclusions (therefore, in conclusion, consequently,...) and contrast (but, except, not, never, no,...), whilst sentiment terminology signals lexical semantic contrast. A comprehensive list of terms is classified according to a scale of sentiment ranging from highly negative to highly positive. The user model covers properties of the user performing the review, and, finally, the domain model specifies the objects and properties that are relevant to the users, for example, properties with binary values (such as has a flash), properties with ranges (such as the number of megapixels, scope of the zoom, or lens size), and multi-slotted properties (such as the warranty).

Wyner further develops the concept of using argument mining as a way to assist manual analysis in Wyner et al. (2015), which describes the development of "Argument Workbench", a tool designed to help the analyst reconstruct arguments from textual sources by highlighting a range of discourse indicators, topics used in the text, domain terminology and speech act terminology. The tool integrates with the DebateGraph software[8], to allow the user to produce detailed argument graphs.

2.5.2 Contextual Clausal Properties

Having considered the argumentative properties intrinsic to a text span, we now move on to look at identifying how a text span is used in the argument

[8] `debategraph.org`

as a whole.

The work of (Moens et al., 2007) on classifying sentences as 'argument' or 'non-argument' is further developed in Palau and Moens (2009), where an additional machine learning technique was implemented to classify each argument sentence as either premise or conclusion, a method referred to as "Argument proposition classification". In this case, the examples considered are extended using material from the European Court of Human Rights (ECHR) and accuracy of classifying sentences as argument increases to 0.80 using the ECHR corpus. Argument proposition classification is carried out using a maximum entropy model and support vector machine, with F-scores of 0.68 for classification as premise and 0.74 for classification as conclusion. Again this work inherits the shortcomings of the earlier research, as the same sentence can be a premise in one context and a conclusion in another.

Such contextual restrictions can however also be an advantage, allowing for example, comments on an article to be connected to the original article based on their relation to it (Aker et al., 2015; Barker and Gaizauskas, 2016). For example, the work of the IBM Debater project in *context dependent evidence detection*, which automatically detects evidence in Wikipedia articles supporting a given claim (Rinott et al., 2015).

In Boltužić and Šnajder (2014) *argument-based opinion mining* is used to determine the arguments on which the users base their opinions. This builds upon previous work in Opinion Mining (as discussed in Section 2.1.1), to include not just the general opinion or stance towards a given topic, but also the arguments on which that stance is based. This is carried out on a specially created corpus of user comments, manually annotated with arguments, using a classifier to predict the correct label from the set of five possible labels (as shown in Table 2.1). The model uses textual entailment and semantic textual similarity features with the best models outperforming the baselines and giving a 0.71 to 0.82 micro-averaged F-score. Although these results give a promising indication of the ability to determine how a comment relates to the argument being made, the topics studied are limited and the training data taken from `procon.org` and `idebate.org` may not be available for, or transfer to, other

topics.

Label	Description: Comment...
A	...explicitly **attacks** the argument
a	...vaguely/implicitly **attacks** the argument
N	...makes **no use** of the argument
s	...vaguely/implicitly **supports** the argument
S	...explicitly **supports** the argument

Table 2.1: Labels for comment-argument pairs (Boltužić and Šnajder, 2014)

The ability to identify even such basic contextual properties offers the opportunity to inform the user and aid in both writing and understanding text. This is again illustrated in Stab and Gurevych (2014b), which aims to identify argument in essays and works towards the long term goal of integrating argumentation classifiers into writing environments. Two classifiers are described. Firstly, for identifying argument components, a multiclass classification is carried out with each clause classified as major claim, claim, premise or non-argumentative. This classifier is trained on a range of feature types, structural features (for example the location and punctuation of the argument component), lexical features (n-grams, verbs, adverbs and modals), syntactic features, discourse indicators and contextual features. Once the argument components have been identified, a second classifier is used to identify argumentative relations (support or non-support). The features used are similar to those for classifying the components, but look at the pairings of clauses. The presented approach achieves 88.1% of human performance for identifying argument components and 90.5% for identifying argumentative relations.

This work is further developed in Nguyen and Litman (2015), where the same methodology and dataset is used, but a Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic model is first generated to separate argument and domain keywords. The output from the LDA algorithm is then post-processed using a minimal seeding of predefined argumentative words to determine argument and domain topics. The same features as (Stab and

Gurevych, 2014b) are then used, replacing n-grams with unigrams of argument words, and numbers of argument and domain words. Using this updated feature set, the accuracy is improved for all of the argument component types: MajorClaim (from 0.48 to 0.59), Claim (from 0.49 to 0.56), and Premise (from 0.86 to 0.88). Whilst these results are promising the relatively low numbers still highlight the difficulties in distinguishing between Claim and MajorClaim, due to the largely context dependent distinction between the two.

The categories from another theory of argumentation structure due to Toulmin (1958), of Data, Claim and Warrant, are similarly difficult to distinguish. Indeed the theoretical impossibility of completely acontextual identification was explored from first principles by Freeman (1991), who showed that under the appropriate circumstances, the difference between Data and Warrant dissolves. With appropriate context, however, the distinction becomes operationally important and was the driver for the first shard task in argument mining, conducted at SEMEVAL2018 by Habernal et al. (2018). The Argument Reasoning Comprehension Task required systems to use a given premise and conclusion to distinguish between two given alternative potential warrants (there is further contextual information available too, with explicitly identified topic and background). E.g.:

> **Topic:** There She Is, Miss America
>
> **Additional info:** In 1968, feminists gathered in Atlantic City to protest the Miss America pageant, calling it racist and sexist. Is this beauty contest bad for women?)
>
> **Argument:** Miss America gives honors and education scholarships. And since ..., Miss America is good for women.
>
> a) scholarships would give women a chance to study
> b) scholarships would take women from the home

The system should in this example choose option (a). Human performance

(following brief training) on this task is at 0.91; system performance in the task varied, with a variety of techniques performing at between 0.50 and 0.70 F-score. Whilst these results seem extremely encouraging, Niven and Kao (2019) suggest that this result is entirely accounted for by exploitation of spurious statistical cues in the dataset, and that by eliminating the major source of these cues, the maximum performance fell from just three points below the average untrained human baseline to essentially random. Niven and Kao counter these effects by the addition of adversarial examples, obtained by negating the claim and inverting the label for each datapoint.

Although the goal of argument mining is the extraction of argumentative structure from natural text, the availability of large quantities of appropriately annotated training data makes this challenging to carry out. An alternative starting point is presented in Peldszus (2014), where a corpus of "microtexts", short texts with explicit argumentation, and little argumentatively irrelevant material is created. The representation of the argument structure within these microtexts is based on Freeman's theory of argumentation structure (Freeman, 1991, 2011), and is viewed as a hypothetical dialectical exchange between a proponent, who presents and defends his claims, and an opponent, who critically questions them. These moves can then be represented as an argument graph, with the nodes representing the propositions expressed in text segments and the edges between them representing different supporting and attacking moves. An agreement between untrained annotators is presented in Peldszus and Stede (2013b). The annotators achieved moderate agreement for certain aspects of the argument graph (e.g. κ=0.52 in distinguishing proponent and opponent segments, or κ=0.58 in distinguishing supporting and attacking segments) yet only a marginal agreement of κ=0.38 on the full labelset describing all aspects of the argument graph. A further study using expert annotators produced significantly higher agreement (κ=0.83) on the full labelset.

The annotation process assigns a list of labels to each segment based on different levels. The 'role'-level specifies the dialectical role (proponent or opponent). The 'typegen'-level specifies the general type, i.e. whether the segment presents the central claim (thesis) of the text, supports or attacks

another segment. The 'type'-level additionally specifies the kind of support (normal or example) and the kind of attack (rebutter or undercutter). Peldszus tests a range of classifiers to automatically classify 'role', 'typegen' and 'type'. The results show that an SVM classifier generally performs best on the most complex labels, suggesting that it deals well with the lower frequencies with which these occur. Meanwhile, the Maximum Entropy and Naïve Bayes classifiers perform best on the simpler and more common labels.

Whilst the results on the microtext corpus are encouraging, the artificial nature of its construction means that such results may not generalise well to unrestricted text. However, this corpus does provide a valuable resource for controlled 'laboratory' testing of argument mining techniques.

2.6 Automatic Identification of Relational Properties

In this section we move on from looking at the identification of clausal properties, to the identification of inter-clausal relations. We look first at general argumentative relations, for example premise/conclusion relationships, and then move on to look at the more complex relationships involved in argumentation schemes and dialogical relations.

2.6.1 Identifying General Argumentative Relations

Identifying relations between pairs of propositions is a more complex and nuanced task than identifying the roles that an individual proposition may take. It is one thing to know, for example, that a given proposition is a premise; much more challenging to determine also for which conclusion (or conclusions) it serves as premise. Approaches to identifying these relations either build upon the prior classification of individual clauses, or aim to extract relations directly.

Palau and Moens (2009), build upon their classification of each argument sentence as either premise or conclusion using a Context-Free Grammar (CFG), produced by grouping manually derived rules. This CFG is used to determine

the internal structure of each individual argument. Whilst the accuracy of classifying sentences as argument or non-argument is 0.80 and F-scores of 0.68 and 0.74 for classification as premise and conclusion respectively, for the harder task of determining argument structure, the accuracy achieved is 0.60.

Peldszus (2014) also builds on the initial task of identifying roles of segments in the Microtext corpus by adding a 'combined'-level, showing, for all types, whether a segment's function holds only in combination with that of another segment (combined) or not (simple). The target is specified by a position relative identifier with a numerical offset identifying the targeted segment relative from the position of the current segment. The prefix 'n' states that the proposition of the node itself is the target, while the prefix 'r' states that the relation coming from the node is the target. Again the results for identifying the target of a relation (maximum F-score of 0.45) are lower than for identifying the roles (maximum F-score of 0.85).

This same microtext corpus is used in Peldszus and Stede (2015), which looks at identifying conflict relations by examining the texts for occurrences of counter-considerations (e.g. "Even though...", or "It has been claimed that...however..."), which the author uses to introduce a potential criticism of their argument, before going on to address the issue and so strengthen their point. This identification is carried out by labelling the textual segments as either 'proponent' or 'opponent' using a linear log-loss model, resulting in an F-score of 0.64 for identifying opposition relations between segments.

Whilst the work discussed thus far in this section builds upon previous identification of component roles before identifying relations, Cabrio and Villata (2012) propose an approach to detect arguments and discover their relationships directly by building on existing work in Textual Entailment (Dagan et al., 2006). Textual Entailment (TE) refers to a "directional relation between two textual fragments, termed text (T) and hypothesis (H), respectively". The relation holds whenever the truth of one text fragment follows from another. In this case, the T-H pair is a pair of arguments expressed by two different users in a dialogue on a certain topic and the TE system returns a judgement (entailment or contradiction) on the argument pair.

A dataset of 300 T-H pairs is created using manually selected topics from Debatepedia[9] which provides pre-annotated arguments (pro or con), and following the criteria defined and by the organisers of the Recognizing Textual Entailment (RTE) challenge[10]. Of these 300 T-H pairs, 200 are used to train (100 entailment and 100 contradiction) and 100 to test (50 entailment and 50 contradiction). The pairs collected for the test set concern completely new topics, never seen by the system, and which are provided in their unlabelled form as input.

TE recognition is carried out using EDITS (Edit Distance Textual Entailment Suite)[11]. EDITS implements a distance-based framework which assumes that the probability of an entailment relation between a given T-H pair is inversely proportional to the distance between T and H. The system uses different approaches to distance computation, providing both edit distance algorithms (cost of the edit operations (insert, delete, etc.) to transform T into H) and similarity algorithms. Each algorithm returns a normalised distance score between 0 and 1. During training, distance scores are used to calculate a threshold that separates entailment from contradiction. Of the EDITS configurations which Cabrio and Villata tested, the highest accuracy is obtained using either Word Overlap or Cosine Similarity (0.66 in both cases), with Token Edit Distance performing significantly less well (accuracy=0.53), suggesting that semantic similarity plays a more important role than syntactic similarity (a result backed up by the comparative analysis of (Aker et al., 2017), which also found syntactic features to be the least informative in all of the experimental settings considered). Whilst these numbers are quite low, this is an interesting result, suggesting that the relationship between topics in an argument gives more of a clue as to how the components relate, than does the way in which those components are expressed. This is carried through in several later works which look at relations between topics and semantic similarity between propositions.

Nguyen and Litman (2016) argue that looking at the content of such pair-

[9] http://www.debatepedia.org

[10] http://www.nist.gov/tac/2010/RTE/

[11] http://edits.fbk.eu/

ings to determine relationships does not make full use of the information available. They propose an approach that makes use of contextual features extracted from surrounding sentences of source and target components as well as from general topic information. Experimental results show that using both general topic information and features of surrounding sentences is effective, but that predicting an argumentative relation will benefit most from combining these two sets of features.

The machine learning approaches to argument mining discussed so far in this section have all used supervised learning to perform classification, however unsupervised learning has also been applied to the task. In Lawrence et al. (2014), a Latent Dirichlet Allocation (LDA) topic model is used to determine the topical similarity of consecutive propositions in a piece of text. The intuition is that if a proposition is similar to its predecessor then there exists some argumentative link between them, whereas if there is low similarity between a proposition and its predecessor, the author is going back to address a previously made point and, in this case, the proposition is compared to all those preceding it to determine whether they should be connected. This assumes that the argument is built up as a tree structure in a depth-first manner, where an individual point is pursued fully before returning to address the previous issues. Although the assumption of a tree structure does not hold for all arguments, it is the case for around 95% of the argument analyses contained in AIFdb (Lawrence et al., 2012b), and 80% of arguments in the Consumer Debt Collection Practices (CDCP) corpus as reported by Niculae et al. (2017). Whilst no evidence is given by Niculae et al. supporting the hypothesis of topical relations with manual analysis of the data, the automated results do support the hypothesis, with a precision of 0.72, and recall of 0.77 recorded when comparing the resulting structure to a manual analysis. It should also be noted that what is being identified here is merely that an inference relationship exists between two propositions, with no indication of the directionality of this inference.

This same approach is implemented in Lawrence and Reed (2015), where

the use of LDA topic models is replaced by using WordNet[12] to determine the semantic similarity between propositions. This change is required to overcome the difficulties in generating a topic model when the text being considered is only a short span, such as an online comment or blog post. The results are comparable to those achieved using LDA, with precision of 0.82 and recall of 0.56. In this case the thresholds are adjusted to increase precision at the expense of recall, as the output from this method is combined with a range of other approaches to determine the final structure, and as such the failure of this approach to identify all of the connections can be compensated for by the other techniques.

A similar approach of assuming a relationship between argument components, if they refer to the same concepts or entities, is used by *AFAlpha* (Carstens et al., 2014), which represents customer reviews as trees of arguments, where a child-parent relationship between two sentences is determined if they refer to the same concepts, with the child being the sentence that has been posted later. A sentence is represented as a set of features, including its semantic characteristics such as metadata about the review in which the sentence appears, as well as features based on the sentences syntactic and lexical nature such as occurrences of certain words and phrase types. A feature vector thus represents each pair of sentences and is classified using a model trained on a data set comprised of data taken from the Q&A debating platform, Quaestio-it[13], and IMDB[14].

Carstens and Toni (2015) continue this line of work focusing on the determination of argumentative relations, and foregoing the decision on whether an isolated piece of text is an argument or not. This focus is based on the observation that the relation to other text is exactly what describes the argumentative function of a particular text span. The paper mentions a number of use cases, describing a method of evaluating claims, by giving a gauge of what proportion of a text argues for or against them. Additionally a preliminary

[12] `http://wordnet.princeton.edu/`

[13] `http://www.quaestio-it.com`

[14] `http://www.imdb.com`

corpus of 854 annotated sentence pairs[15] is provided, with each sentence pair labelled with $L \in \{A, S, N\}$, where A = Attack, S = Support, or N = Neither (including both cases where the two sentences are unrelated and those where they are related, but not in an argumentative manner.)

The important role played by similarity is also exploited by Gemechu and Reed (2019) who borrow notions of aspect, target concept and opinion from opinion mining, and use these to decompose ADUs down into finer-grained components, and then use similarity measures between these components to identify argument relations. Such *decompositional argument mining* not only performs well on diverse single-author arguments (outperforming the techniques of Peldszus and Stede on their Microtext corpus, and of Stab and Gurevych on their AAEC corpus) but also on arguments situated in dialogue (albeit at lower levels of performance: F1 ranging from 0.74 to 0.77 on both Microtext and AAEC, and 0.63 on US2016).

Finally, (Wachsmuth et al., 2018) highlights an interesting link between similarity and argumentative relations. The work presented aims to determine the best counterargument to any argument without prior knowledge of the argument's topic. The best performing model tested rewards a high overall similarity between a potential counterargument and the given argument's conclusion and premises whilst punishing those counterarguments that are too similar to either of them. To some extent, this result captures the intuition that argumentative relations occur where something different is being said about the same topic.

2.6.2 Identifying Complex Argumentative Relations

The ability to successfully extract premises and conclusions is built upon in Feng and Hirst (2011), which presents the first step in the long term goal of a method to reconstruct enthymemes, by first, classifying to an argumentation scheme (Walton et al., 2008) then fitting the propositions to the template and finally, inferring the enthymemes. For the first step of fitting one of the top five most commonly occurring argumentation schemes to a pre-determined

[15] Available at `www.doc.ic.ac.uk/~lc1310/`

argument structure, accuracies of 0.63–0.91 are recorded in one-against-others classification and 0.80–0.94 in pairwise classification. As in Moens et al. (2007), the Araucaria corpus is used with complex Argument Units (AUs) first broken into simple AUs (with no embedded AUs). The AUs using the top five most common argumentation schemes are then selected and a classifier trained on both features specific to each individual scheme and a range of general linguistic features, in order to obtain the scheme. Although these results are promising, and suggest that identifying scheme instances is an achievable task, they do rely on the prior identification of premises and conclusions, as well as the basic structure which they represent. Whilst this approach does not identify the roles of individual propositions in the scheme, knowing what type of scheme links a set of propositions is both a useful task in its own right and offers potential for subsequent processing to determine proposition types for each scheme component. This is a substantially easier task once the scheme type is known.

Another approach to identifying the occurrence of schemes is given in Lawrence and Reed (2015), where, rather than considering features of the schemes as a whole, the individual scheme components are identified and then grouped together into a scheme instance. In this case, only two schemes ('Expert Opinion' and 'Positive Consequences') are considered and classifiers trained to identify their individual component premises and conclusion. By considering the features of the individual types of these components, F-scores between 0.75 and 0.93 are given for identifying at least one component part of a scheme.

The approach followed by (Feng and Hirst, 2011) is similar in nature to the first steps suggested by (Walton, 2011), where a six-stage approach to identifying arguments and their schemes is proposed. The first of these stages is the identification of the arguments occurring in a piece of text; this is followed by identification of specific known argumentation schemes. Walton, however, points out that beyond this initial identification there are likely to be issues differentiating between similar schemes and suggests the development of a corpus of borderline cases to address the issue.

As Walton points out, the automatic identification of argumentation schemes remains a major challenge. As discussed in Section 2.2.4, a large number of scheme classifications exist, with additional domain specific schemes utilised in specific areas. For example, as part of the rule-based tool for semi-automatic identification of argumentative sections in text presented in Wyner et al. (2012), a consumer argumentation scheme (listing 2.1) is described and the structure of this scheme used to guide the argument identification process.

Listing 2.1: Consumer Argumentation Scheme

```
Premise: Camera X has property P
Premise: Property P promotes value V for agent A
Conclusion: Agent A should Action1 camera X.
```

Similarly, (Green, 2015) lists ten custom argumentation schemes targeted at genetics research articles. For example, one of the schemes presented, 'Failed to Observe Effect of Hypothesized Cause', looks for situations where specific properties were not observed, and where it is assumed that a specific condition that would result in those properties is present, leading to the conclusion that the condition may not be present. Green (2018a) further argues for schemes expressed in terms of domain concepts rather than by generic definitions as in those of (Walton et al., 2008) carrying out a pilot annotation study of schemes for 15 arguments in the Results/Discussion section of biological/biomedical journal articles. Green (2018b) then explores how argumentation schemes in this domain can be implemented as logic programs in Prolog and used to extract individual arguments. In this case, the schemes are formulated in terms of semantic predicates obtained from a text by use of BioNLP (biomedical/biological natural language processing) tools.

Regardless of the theoretical backdrop, schemes generally introduce as much complexity as they do opportunity from annotation through to automated analysis. To pick an example from a substantially different theoretical apprach, Musi et al. (2016) present a novel set of guidelines for the annotation of argument schemes based on the Argumentum Model of Topics (Rigotti and Morasso, 2010). This framework offers a hierarchical taxonomy of argument schemes based on linguistic criteria which are distinctive and applicable to a

broad range of contexts, aiming to overcome the challenges in annotating a broad range of schemes.

With the data currently available, the ontologically rich information available in argumentation schemes has been demonstrated to be a powerful component of a robust approach to argument mining. Collaboration amongst analysts as well as the further development of tools supporting argumentation schemes is essential to growing the datasets required to improve on these techniques. Clear annotation guidelines and the development of custom argumentation schemes for specific domains, will hopefully result in a rapid growth in the material available and further increase the effectiveness of schematic classification.

Dialogical Relations

Whilst some of the previously mentioned argument mining techniques have worked with data that is dialogical in nature, such as user comments and online discussion forums, none of these have focused on using the unique features of dialogue to aid in the automatic analysis process, producing an analysis that captures both the argumentative and dialogical structure. For example although (Pallotta et al., 2004; Rienks et al., 2005) consider dialogical data, in both cases they do not consider the specific dialogical relations between utterances.

Similarly there is a large body of work studying the nature of dialogue both in terms of dialogue modeling, which captures the nature and rules of a dialogue, and dialogue management, which takes a more participant oriented viewpoint in determining what dialogical moves to make (Traum, 2017). However, there is currently little work that puts these models to work in enhancing argument mining techniques. It seems clear that by modeling a dialogue and understanding that the next move a participant is likely to make will be 'disagreeing,' for example, we would be able to obtain the argumentative structure easily. In this section we discuss formalisations of dialogue protocols and then move on to cover the work that has been done to apply this knowledge to argument mining.

In the case of more formally structured dialogues, a protocol for the dialogue can be described, and specified in a language such as the FIPA Agent Coordination Language (McBurney and Parsons, 2009), the Dialogue Game Description Language (DGDL) (Bex et al., 2014a) or the Lightweight Coordination Calculus (Robertson, 2004). Such dialogue games have been developed to capture a range of more structured conversations, for example, to facilitate the generation of mathematical proofs (Pease et al., 2017) or help reach agreement on which course of action to take in specific circumstances (Atkinson et al., 2005). In these cases, software such as Arvina (Lawrence et al., 2012a) or D-BAS (Krauthoff et al., 2018) can be used to both run the dialogue according to the specified rules and automatically capture the argumentative structure generated as the dialogue progresses. These structures can then be used to allow for mixed initiative argumentation (Snaith et al., 2010), where a combination of human users and software agents representing the arguments made by other people can take part in the same conversation, using retrieval-based methods to select the most relevant response (Le et al., 2018). In such scenarios, the contributions of human participants can be interpreted by virtue of their dialogical connections to the discourse, allowing a small step towards mining argument structure from natural language.

Although formally structured dialogues can be captured and exploited in this way, many real world dialogues follow only very limited rules and the challenge of identifying the argumentative structure in free form discussion is complex. However, even very informal dialogues nevertheless provide additional data beyond that available in monologue, which can be used to help constrain the task.

Amongst other such features, Budzynska et al. (2014) identify illocutionary forces and dialogue transitions. Illocutionary forces are the speech act type (Austin, 1962) of utterances. Their automatic recognition in Illocutionary Structure Parsing (Budzynska et al., 2016) is similar to Dialogue Act Annotation (Bunt et al., 2010) though often rather more specific. Automatic distinction between rhetorical, pure, and 'assertive' questioning, for example is nuanced and challenging. The preliminary results reported in Budzynska

et al. (2016) point to accuracy of 78% on this task, but the datasets used are very small (n = 153).

Al Khatib et al. (2018) identify six distinct 'discourse acts' ('Socializing', 'Providing evidence', 'Enhancing the understanding', 'Recommending an act', 'Asking a question', and 'Finalizing the discussion') in deliberative discussions. As a first step towards determining the best possible move for a participant in a deliberative discussion, Al Khatib et al. train an SVM model to classify examples of these discourse acts from Wikipedia data. Whilst the classifier achieves low F-scores for 'socializing', 'recommending an act', and 'asking a question', these are the categories with the smallest number of examples in the dataset to draw from – 83, 137 and 106 turns respectively. Performance on those acts with more examples is much better: 'Providing evidence' (781 turns, F-score = 0.69), 'Enhancing the understanding' (671 turns, F-score = 0.58), and 'Finalizing the discussion' (622 turns, F-score = 0.71). These results are encouraging and suggest that with more data, further improvements could be expected.

Dialogue transitions, on the other hand, connect together dialogical moves. In Inference Anchoring Theory (IAT) (Budzynska et al., 2014), illocutionary connections are anchored in these transitions. This explicit connectivity can be used to handle complex phenomena such as indexicality (where the propositional content of one locution can only be reconstructed by reference to another locution, for example: "Isn't that a source of injustice?" – "Definitely not."). Budzynska et al. suggest that the patterns provided by transitions can constrain the mining process by defining expectations (for example, if an assertive question is followed by a negative polarity indexical assertion then such a transition anchors the illocutionary connection of disagreeing). There are no results yet reported testing this hypothesis.

2.7 Conclusion

Argument mining techniques have been successfully developed to extract details of the argumentative structure expressed within a piece of text, focusing

on different levels of argumentative complexity as the domain and task require. For each task, we have considered work carried out using a broad range of techniques, including statistical and linguistic methods. We have presented a hierarchy of task types based on increasing argumentative complexity. First looking at the identification of argument components and the determination of their boundaries, we have then moved on to consider the role of individual clauses (both intrinsic, such as whether the clause is reported speech, and contextual such as whether the clause is the conclusion to an argument). Finally, we have considered the identification of a range of argumentative relations from simple premise/conclusion relationships, to whether a set of clauses form and instance of an argumentation scheme.

There fact remains that argument mining is a difficult task; as Moens (2018) points out, "a lot of content is not expressed explicitly but resides in the mind of communicator and audience". It seems that to overcome this challenge we need to look at the broader picture in which argument occurs. In this regard, works which either takes a more holistic 'end-to-end' view Stab and Gurevych (2017); Persing and Ng (2016); Potash et al. (2017), or which aim to harness external data sources Rinott et al. (2015); Lawrence and Reed (2017a), seem to point the way.

The success of these techniques and the development of techniques for analysing dialogical argument, offers hope that techniques can be developed for automatically identifying complex illocutionary structures and the argumentative structures they build. We have also seen how these techniques can be combined, tying together statistical identification of basic structure, linguistic markers and identifying scheme components. In so doing, the resulting argument structures offer a more complete analysis of the text than any of these methods provide on their own.

Argument mining remains profoundly challenging, and traditional methods on their own seem to need to be complemented by stronger, knowledge-driven analysis and processing. However, the pieces required to successfully automate the process of turning unstructured data into structured argument are starting to take shape. As the volume of analysed argument continues to increase, and

existing techniques are further developed and brought together, rapid progress can be expected.

Chapter 3

Argument Data

This chapter explores argument data, looking first at the most widely used argument corpora and considering their specific strengths and weaknesses, before then moving on to describe the specific datasets used throughout the remainder of this thesis.

3.1 Argument Corpora

One of the challenges faced by current approaches to argument mining is the lack of large quantities of appropriately annotated arguments to serve as training and test data. Several recent efforts have been made to improve this situation by the creation of corpora and argumentative datasets across a range of different domains. These efforts can be broken down into two main categories: manually annotated corpora of argumentative components and structure found in natural language text; and corpora of pre-structured text where the argumentative structure is captured as part of its creation, or the argumentative structure can be inferred from other existing structural features.

3.1.1 Manually Annotated Corpora

The Internet Argument Corpus (IAC) (Walker et al., 2012) is a corpus for research in political debate on internet forums. It consists of ∼11,000 discussions, ∼390,000 posts, and some ∼73,000,000 words. Subsets of the data have been annotated for topic, stance, agreement, sarcasm, and nastiness among others.

The IAC is further developed in the Internet Argument Corpus (IAC) version 2 (Abbott et al., 2016), a collection of corpora for research in political debate on internet forums. It consists of three datasets: 4forums (414K posts), ConvinceMe (65K posts), and a sample from CreateDebate (3K posts). It includes topic annotations, response characterizations (4forums), and stance, though argument annotation in both IAC datasets is rather limited by comparison to that available in other datasets.

Whilst providing the largest corpus of annotated argumentation, the IAC is limited by the sparsity of annotations it contains, with large sections of the corpus only being labelled as "argumentative". There have been many attempts to provide a more detailed and comprehensive coverage of all the arguments which a text contains, in most cases starting with a more limited selection of text from a single specific domain. For example, Green (2014) aim to create a freely available corpus of open-access, full-text scientific articles from the biomedical genetics research literature, annotated to support argument mining research. However, there are challenges to creating such corpora, such as the extensive use of biological, chemical, and clinical terminology in the BioNLP domain requiring annotators trained in the field. These challenges are highlighted in Green (2015), where preliminary work on guidelines for the manual identification of ten custom argumentation schemes targeted at genetics research articles, is presented. For example, one of the schemes presented, 'Failed to Observe Effect of Hypothesized Cause', looks for situations where specific properties were not observed, and where it is assumed that a specific condition that would result in those properties is present, leading to the conclusion that the condition may not be present. Twenty-three students were assessed on their ability to identify instances of these schemes after having read the guidelines, and the results show a mean accuracy of only 49%. It can be seen from these results that the classification of such nuanced argument schemes is not a straightforward task. This suggests the need for both more rigorous scheme definitions, with particular attention given to error analysis of those schemes which are commonly confused, as well as the development of annotation guidelines taking these issues into account.

The Argument Annotated Essays Corpus (AAEC) (Stab and Gurevych, 2014a) consists of argument annotated persuasive essays, and features topic and stance identification, annotation of argument components, and argumentative relations. Drawn from 90 English language essays posted in the *writing feedback* section of the website essayforum[1]. The final corpus contains 90 major claims, 429 claims, and 1,033 premises, connected by 1,312 support and 161 attack relations. The AAEC version 2 (Stab and Gurevych, 2017) extends this to 402 essays, with 751 major claims, 1,506 claims, and 3,832 premises, connected by 3,613 support and 219 attack relations.

A random sample of 102 essays taken from the AAEC have been further annotated, as described in Carlile et al. (2018), to also include a persuasiveness score for each argument as well as scores for attributes that potentially impact persuasiveness (Eloquence, Specificity, Relevance, and Evidence), the means of persuasion (Ethos, Pathos or Logos) and the types of both claims and premises. This addition to AAEC has already shown potential in developing automated persuasiveness scoring for essays (Ke et al., 2018).

Kirschner et al. (2015) present a corpus of twenty-four German language articles were selected from the education research domain, and annotated using a custom designed tool (DiGAT). The annotation scheme used identifies binary relations between argument components, which in this work correspond to sentences from the original texts. Four types of relation are identified: 'support', 'attack', 'detail' and 'sequence'. The first two of these relations are argumentative, whereas the latter two are discourse relations similar to the 'sequence' and 'background' relations of Rhetorical Structure Theory (RST) (Mann and Thompson, 1987). The results of annotation using this scheme are represented as graph structures, and a range of methods to determine inter annotator agreement for these structures are considered. Despite the complexity of the articles being analysed, the results show multi-κ[2] values up to 0.63. Whilst this result is fair for such a complex annotation task, several specific areas are identified which reduce agreement. Similar categories were particularly problematic, for example, in many cases disagreement was due to

[1] `http://www.essayforum.com`

[2] An extension of Cohen's Kappa allowing for multiple annotators (Hubert, 1977).

confusion between 'support' and 'detail' or 'support' and 'sequence relation'. Although these differences could potentially be improved by more detailed annotation guidelines, the authors argue that in many cases several correct solutions exist, with both labellings being correct.

Legal texts are the focus of Walker et al. (2014), where a type system is developed for marking up successful and unsuccessful patterns of argument in U.S. judicial decisions. Building on a corpus of vaccine-injury compensation cases that report factfinding about causation, based on both scientific and non-scientific evidence and reasoning, patterns of reasoning are identified and used to illustrate the difficulty of developing a type or annotation system for characterising these patterns. A further example of legal material is the ECHR corpus (Mochales and Ieven, 2009), a set of documents extracted from legal texts of the European Court of Human Rights (ECHR). The ECHR material, whilst not annotated specifically for argumentative content, contains a standard type of reasoning and structure of argumentation which means that the corpus can be easily adapted to serve as data for argument mining.

A different domain is considered in Kiesel et al. (2015), which presents a corpus of 200 newspaper editorials annotated for their argumentative structure. The annotation is based on a model consisting of explicit argumentative units, and the implicit argumentative relations (i.e. support or attack) between them. In this case, an argumentative unit is understood to be a segment of the original text containing at least one proposition. Argumentative relations are considered as the links from one unit to the unit that it most directly supports or attacks.

Such efforts add to the volume of currently available data for which at least some elements of the argumentative structure have been identified. The most comprehensive and completely annotated existing collection of such data is the openly accessible database, AIFdb[3] (Lawrence et al., 2012b), containing over 18,000 Argument Interchange Format (AIF) (Chesñevar et al., 2006) argument maps, with over 2.1m words and 200,000 claims in fourteen different languages[4]. These numbers are growing rapidly, thanks to both the increase

[3] `http://www.aifdb.org`

[4] Amharic, Chinese, Dutch, English, French, German, Hindi, Italian, Japanese, Polish,

in analysis tools interacting directly with AIFdb and the ability to import analyses produced with the Rationale and Carneades tools (Bex et al., 2012). Indeed, AIFdb aims to provide researchers with a facility to store large quantities of argument data in a uniform way. AIFdb web services allow data to be imported and exported in a range of formats to encourage collaboration between researchers independent of the specific tools and data format that they require.

Additionally, several online tools such as DebateGraph[5], TruthMapping[6], Debatepedia[7], Agora[8], Argunet[9] and Rationale Online[10] allow users to create and share argument analyses. Although these tools are helping to increase the volume of analysed argumentation, they generally do not offer the ability to access this data and each use their own formats for its annotation and storage. In order to help overcome this challenge, AIFdb offers the facility to import and convert Rationale and Carneades analyses into AIF. At the moment, though, many research projects continue to introduce ad hoc, idiosyncratic data representation languages for argumentation and debate, which can limit reusability, integration and longevity of the datasets.

3.1.2 Pre-structured Argument Corpora

Whilst the previously discussed datasets can be viewed as "fully" structured argument data, there is an increasing usage of larger "semi-structured" argumentative data sources, from which argumentative data can be extracted. The most striking example of such are recent datasets gathered from the "ChangeMyView" (CMV) Reddit subcommunity[11] (Tan et al., 2016; Hidey and McKeown, 2018). This data takes the form of discussion threads where the original poster of a thread provides a viewpoint on a specific topic, and other users re-

Portuguese, Russian, Spanish and Ukrainian

[5] `http://debategraph.org`

[6] `https://www.truthmapping.com`

[7] `http://www.debatepedia.org`

[8] `http://agora.gatech.edu/`

[9] `http://www.argunet.org/`

[10] `https://www.rationaleonline.com/`

[11] `https://www.reddit.com/r/changemyview/`

ply with comments aiming to change this view. If the original poster finds that a comment succeeds in changing their viewpoint, they can reply with a 'delta' symbol indicating this. Whilst this data is not strictly argumentative, there are strong indicators of argumentative structure: direct responses, for example, often include counterarguments to the original post. Indeed, Hua and Wang (2017) uses CMV data to both train and evaluate a model for automatically generating arguments of the opposing stance for a given statement.

In addition to these corpora of structured argument data, there are large corpora of unstructured data available that are rich in argumentative structure, from, for example, Wikipedia, Twitter, Google Books, and product reviews from websites such as Amazon and epinions.com. Whilst these corpora may be useful for certain argument mining techniques, such as those using unsupervised learning methods, there are limits on their utility imposed, inevitably, by their lack of annotation.

Despite the lack of marked argument structure, Wikipedia, in particular, represents a considerable amount of data rich in argumentative content. In Aharoni et al. (2014), work towards annotating articles from Wikipedia using a meticulously monitored manual annotation process is discussed. The result is a corpus of 2,683 argument elements, collected in the context of 33 pre-defined controversial topics, and organised under a simple structure detailing a claim and its associated supporting evidence.

In their far-ranging work on Project Debater[12], IBM have made extensive use of Wikipedia and other data to create the first AI system that can debate humans on complex topics. Debater can respond to a given topic by automatically constructing a set of relevant pro/con arguments phrased in natural language. For example, when asked for responses to the topic "The sale of violent video games to minors should be banned", an early prototype of Debater scanned approximately 4 million Wikipedia articles and determined the ten most relevant articles, scanned all 3,000 sentences in those articles, detected sentences which contain candidate claims, assessed their pro and con polarity and then presented three relevant pro and con arguments[13], with more recent

[12] `https://www.research.ibm.com/artificial-intelligence/project-debater/`

[13] `http://www.kurzweilai.net/introducing-a-new-feature-of-ibms-watson-the-debater`

developments also working towards choosing the most convincing of these arguments (Gleize et al., 2019), expanding the topic of the debate (Bar-Haim et al., 2019), and providing "first principle" debate points, commonplace arguments which are relevant to many topics, where specific data is lacking (Bilu et al., 2019). These abilities are the result of ongoing work to extract meaningful argument data from large corpora. In Levy et al. (2014), the challenge of detecting *Context Dependent Claims (CDCs)* in Wikipedia articles was first addressed, showing how, given a topic and a selection of relevant articles, a selection of "general, concise statements that directly support or contest the given topic" can be found. This work was followed in Rinott et al. (2015) where extracting supporting evidence from Wikipedia data for a given CDC was addressed. Bar-Haim et al. (2017) introduced the task of claim stance classification, that is, detecting the target of a given CDC, and determining the stance towards that target. Levy et al. (2017) further developed CDC identification, removing the need for pre-selected relevant articles, by first deriving a *claim sentence query* to retrieve CDCs from a large unlabelled corpus. (Indeed, this retrieval task is increasingly becoming a distinct and challenging task in its own right, with applications such as *args.me* (Wachsmuth et al., 2017a) and new shared tasks such as Touche[14] driving the area forward). Such large volumes of CDCs can be used both as potential points to be made by the debater system as well as to aid in the interpretation of spoken material containing breaks, repetitions, or other irregularities (Lavee et al., 2019). The method introduced by Levy et al. is used in Shnarch et al. (2018) to generate *weakly labelled data* (data of low quality compared to manual annotation, but which can be automatically obtained in large quantities) and then combined with a smaller quantity of high quality, manually labelled data (*strongly labelled data*). Using the combined strongly and weakly labelled dataset as training data resulted in improved performance for topic-dependent evidence detection, suggesting that this kind of data gathering can be a valuable asset, particularly in data-hungry neural network systems. The annotated datasets used in this and other Project Debater work are all available online[15].

[14] `http://touche.webis.de`

[15] `http://www.research.ibm.com/haifa/dept/vst/debating_data.shtml`

Bosc et al. (2016) address another rich online data source, taking data from Twitter and defining guidelines to detect 'tweet-arguments' among a stream of tweets about a certain topic, before then pairing the identified arguments, and finally, providing a methodology to identify which kind of relation holds between the arguments composing a pair, i.e., support or attack. Bosc, Cabrio, and Villata report agreement of α[16] of 0.81 for detecting argumentative tweets, and α of 0.67 for argument linking, with the resulting DART (Dataset of Arguments and their Relations on Twitter) dataset containing 4,000 tweets annotated as argument/not-argument with 446 support and 122 attack relations.

In Houngbo and Mercer (2014), a straightforward feature of co-referring text – presence of the lexeme, "this" – is used to build a self-annotating corpus extracted from a large biomedical research paper dataset. This is achieved by collecting pairs of sequential sentences where the second sentence begins with "This method...", "This result...", or "This conclusion...", and then categorising the first sentence in each pair respectively as Method, Result or Conclusion sentences. In order to remove outliers in the dataset, a multinomial Naïve Bayes classifier was trained on the collected Method, Result or Conclusion sentences, and sentences from this set that were then classified to the same category with less than 98% confidence were removed. This reduced corpus was then used as training data to identify Method, Result and Conclusion sentences using both SVM and Naïve Bayes classifiers. These classifiers show an average F-score of 0.97 with Naïve Bayes and 0.99 with SVM, and are further tested on the corpus used by Agarwal and Yu (2009) where sentences are classified in the same way. By using this approach Houngbo and Mercer are able to improve on the results from Agarwal and Yu whose results show an F-score of 0.92 using 10-fold cross-validation. Despite the limited nature of this task, only identifying specific types of sentence and not giving any idea of the relations between them, these results show that by extending the training data available, substantial improvements in classifying sentences can be made.

Lawrence and Reed (2017a) take a similar approach to Houngbo and Mer-

[16]Krippendorff's alpha (α) is a reliability coefficient developed to measure the agreement among coders (Krippendorff, 1980)

Name	Description	Size	IAA	URL	Reference
AIFdb Corpora					
Argumentation Schemes	Examples of occurrences of Walton's argumentation schemes found in episodes of the BBC Moral Maze Radio 4 programme.	6,704 words	Single annotator	http://corpora.aifdb.org/schemes	(Lawrence and Reed, 2016)
Digging By Debating	Collection of analyses of 19th century philosophical texts from the Hathi Trust collection.	35,789 words	Single annotator	http://corpora.aifdb.org/dbyd	(Murdock et al., 2017)
Dispute Mediation	Argument maps of mediation session transcripts.	26,923 words	$\kappa = 0.68$	http://corpora.aifdb.org/mediation	(Janier and Reed, 2016)
MM2012	Analyses of all episodes from the 2012 summer season of the BBC Moral Maze Radio 4 programme.	29,068 words	$\kappa = 0.55$ (types), 0.61 (relations)	http://corpora.aifdb.org/mm2012	(Budzynska et al., 2014)
US2016	2016 US presidential elections: annotations of selected excerpts of primary and general election debates, combined with annotations of selected excerpts of corresponding Reddit comments.	87,064 words	$\kappa = 0.75$	http://corpora.aifdb.org/US2016	(Visser et al., 2018b)
Imported into AIFdb					
AraucariaDB	An import of 661 argument analyses produced using Araucaria and stored in the Araucaria database.	62,881 words	Single annotator	http://corpora.aifdb.org/araucaria	(Reed, 2006)
AraucariaDBpl	A selection of over 50 Polish language analyses created using the Polish version of Araucaria.	2,654 words	Single annotator	http://corpora.aifdb.org/araucariapl	(Budzynska, 2011)
eRulemaking	Argument maps of 67 comment threads from regulationroom.org.	26,083 words	$\kappa = 0.73$	http://corpora.aifdb.org/RRD	(Park and Cardie, 2014)
Internet Argument Corpus (IAC)	Consisting of 11,000 discussions and developed for research in political debate on internet forums. Subsets of the data have been annotated for topic, stance, agreement, sarcasm, and nastiness among others.	1,031,398 words	$\kappa = 0.22$-0.60, $\bar{\kappa} \approx 0.47$	http://corpora.aifdb.org/IAC	(Walker et al., 2012)
Language Of Opposition	Used in Rutgers for the SALTS project (`http://salts.rutgers.edu/`).	48,666 words	Not reported	http://corpora.aifdb.org/looc1	(Ghosh et al., 2014)
Microtext	112 manually created, short texts with explicit argumentation, and little argumentatively irrelevant material.	7,828 words	$\kappa = 0.83$	http://corpora.aifdb.org/Microtext	(Peldszus, 2014)
Available elsewhere					
Argument Annotated Essays	The corpus consists of argument annotated persuasive essays including annotations of argument components and argumentative relations.	147,271 words	$\kappa = 0.64$-0.88 (types), 0.71-0.74 (relations)	https://bit.ly/2OlRZnt	(Stab and Gurevych, 2017)
Argument Annotated User-Generated Web Discourse	User comments, forum posts, blogs and newspaper articles annotated with an argument scheme based on an extended Toulmin model	84,673 words	$\alpha_U = 0.51$-0.80	https://bit.ly/2vdkHOD	(Habernal and Gurevych, 2017)
Consumer Debt Collection Practices (CDCP)	User comments about rule proposals by the Consumer Financial Protection Bureau collected from an eRulemaking website	∼88,000 words	$\alpha = 0.65$ (types), 0.44 (relations)	http://joonsuk.org	(Niculae et al., 2017)
Internet Argument Corpus (IAC) 2	Corpus for research in political debate on internet forums. It includes topic annotations, response characterizations, and stance.	∼500,000 forum posts	Not reported	https://nlds.soe.ucsc.edu/iac2	(Abbott et al., 2016)
IBM Project Debater Datasets	Collection of annotated data sets developed as part of Project Debater to facilitate this research. Organized by research sub-fields.	Various	Various	https://ibm.co/2OlqieA	(Rinott et al., 2015), (Levy et al., 2017) etc

Table 3.1: Significant argumentation datasets available online

cer, using 'discourse indicators' (connectives such as "because", "however" etc.) in place of "this". In this work, the topic of a given text is first identified and a web search carried out to retrieve related documents. Sentences containing discourse indicators showing support relations are then found within the retrieved documents and these sentences are split either side of the indicator to give possible premise conclusion pairs. Despite this being a noisy dataset, with potential off-topic sentences and cases where the indicator has been used for a different reason, it is shown that a topic model can be built from large numbers of these pairs resulting in stereotypical patterns of support on the given topic.

Similarly, Habernal and Gurevych (2015), use large volumes of unlabeled data from online debate portals. By identifying clusters of both sentences and posts from these debate portals which contain similar phrases, and then finding the centroids of these clusters, 'prototypical arguments' are identified. Al-Khatib et al. (2016) likewise leverage online debate portals, generating annotations by automatically mapping source data, in this case the labelled text components from the idebate.org (e.g. 'Introduction', 'point', 'counterpoint'), to a set of predefined class labels to create a large corpus with argumentative and non-argumentative text segments from several domains.

An alternative approach to generating argument corpora is presented in Peldszus (2014), where a corpus of "microtexts", originally produced in German and then also professionally translated into English, is created. These texts, were generated by asking participants to write a text, approximately five segments long in which: all segments are argumentatively relevant; there is a segment acting as the main claim of the text; all other segments are supporting/attacking the main claim or another segment; at least one possible objection to the claim is considered in the text. Whilst the this method of generating argument data produces very clear examples of argumentative relations, the artificial nature of its construction means that results obtained on the dataset may not generalise well to unrestricted text. However, this corpus does provide a valuable resource for controlled 'laboratory' testing of argument mining techniques. Further details of the microtext corpus, as well as the other

corpora discussed in this section can be seen in Table 3.1.

3.2 Experimental Dataset

Whilst there is a broad, and growing, range of argumentative datasets available for use in argument mining, many of these suffer from either a lack of comprehensive coverage of the data, coarse-grained analysis of the argument structure, or limited applicability to real-world arguments. The experiments described in the remainder of this thesis use the 'US2016G1tv' corpus (Visser et al., 2020a): a challenging real-world argumentative analysis of the first 2016 United States of America presidential election general debate between Hillary Clinton and Donald Trump (26 September 2016, Hempstead, NY). The US2016D1tv (first televised Democratic Primary debate) and US2016R1tv (Republican) corpora are also used to provide additional training data in some cases[17].

Transcripts of these debates were annotated using the OVA analysis tool (Lawrence et al., 2017b), stored in AIFdb (Lawrence et al., 2012b), and collected in AIFdb Copora (Lawrence and Reed, 2014). The US2016G1tv corpus is freely available online at `http://corpora.aifdb.org/US2016G1tv`.

Annotation was performed on the basis of Inference Anchoring Theory (IAT) (Budzynska and Reed, 2011). IAT builds on insights from discourse and conversation analysis, speech act theory, and argumentation studies, as a way of explaining how the propositional reasoning that is appealed to in argumentation is anchored in discourse (whether written or spoken). IAT annotation results in an Argument Interchange Format (AIF) (Chesñevar et al., 2006) compliant graph representation of both the reconstructed argumentation structure and its discursive anchoring in the analysed text segments.

IAT underpins the annotation guidelines used by the four expert annotators involved in the annotation of the US2016G1tv corpus. Based on a 11.3% sample annotated by two annotators, the agreement between the annotators was substantial (according to the Landis and Koch (1977) interpretation), with a Cohen's κ (Cohen, 1960) of 0.61. Duthie et al. (2016b) have, how-

[17]These corpora combine to give the US2016tv corpus, which along with corresponding social media reactions (US2016reddit) comprise the US2016 corpus

ever, argued that Cohen's κ misrepresents the interdependency between some of the sub-tasks involved in the annotation process. For example, a difference in initial segmentation of the text can then have cascading effects on the results for structuring the segments. To do justice to such interdependency, Duthie et al. propose to calculate a Combined Argument Similarity Score (or CASS-κ) by combining independent agreement scores for the sub-tasks of text segmentation, discourse annotation, and propositional annotation. When taking into account the interplay between these constitutive tasks, the average inter-annotator agreement in terms of CASS-κ is 0.752.

While the full annotation guidelines[18] deal with complex issues such as anaphoric references, epistemic modalities, repetition, punctuation, discourse indicators, interposed text, and reported speech, we summarise below those aspects of the annotation that are essential for a proper understanding of the corpus study.

Locutions: The original text is first segmented into locutions. A locution consists of a speaker designation and an 'argumentative discourse unit' (ADU) (Peldszus and Stede, 2013a), a text span with discrete argumentative function (often directly resulting in the introduction of an inference, conflict or rephrase in the argumentation structure – see below). In accordance with the AIF ontology, locutions are modelled as L-nodes, a sub-type of I-node. It should be noted that the techniques presented in the remainder of this thesis do not address the segmentation task, instead starting with manually segmented text and viewing segmentation as a separate challenge[19].

Transitions: Functional discourse relationships are represented as transitions connecting the segmented locutions. The transitions reflect the dialogue protocol underpinning the discourse. Transitions, or TA-nodes, are a type of S-node that connects L-nodes.

Illocutionary connections: The communicative intention encapsulated in a locution is annotated by means of illocutionary connections that relate the locutionary to the propositional dimension of the analysis. In AIF terms,

[18] `http://arg.tech/US2016-guidelines`

[19] This is in line with almost all other argument mining work which begins with either manually pre-segmented text, or simply segments by sentence boundary

illocutionary connections are YA-nodes, a sub-type of S-node.

Propositions: Most illocutionary connections lead to the reconstruction of the propositional content of the associated locution. Propositions are modelled as I-nodes.

Inference, conflict and rephrase: Generally connecting one proposition to another, the argumentative relations of inference, conflict and rephrase respectively indicate justificatory defence, refutatory incompatibility, and revisionary reformulation. The propositional relations are modelled as sub-types of S-nodes: as RA-, CA-, and MA-nodes. Walton's argumentation schemes (Walton, 1996) have been developed in full for the AIF as types for RA-nodes (Rahwan et al., 2007) and, using these, the US2016G1tv corpus has been extended with full argumentation scheme annotation (see Chapter 8) making this corpus the largest collection of annotated scheme instances (replacing the Araucaria corpus used by (Feng and Hirst, 2011)).

Table 3.2: Proposition and propositional relation counts for the US2016tv corpora

Corpus	Propositions	Inference	Conflict	Rephrase
US2016G1tv	1473	505	79	140
US2016R1tv	1368	482	61	88
US2016D1tv	1439	564	54	105

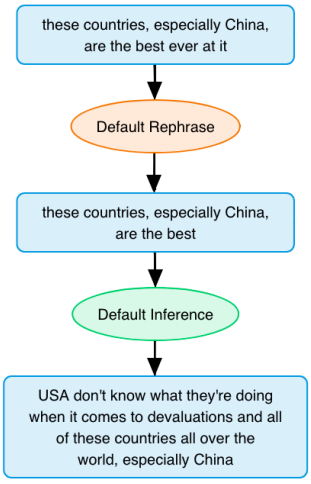


Figure 3.1: An example of a rephrase (MA) relation in US2016G1tv.

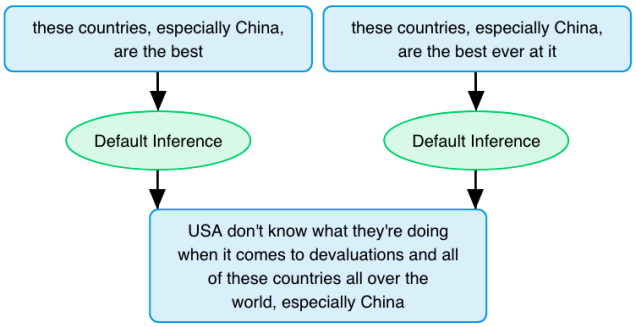


Figure 3.2: An example of a rephrase (MA) relation mapped to two separate inference relations.

Table 3.2, shows the most relevant properties of the 17,190-word (tokens) US2016G1tv corpus, as well as those for US2016D1tv and US2016R1tv[20].

Two aspects of this annotation require further consideration when used for the purposes of argument mining: rephrase relations and reported speech. Figure 3.1 shows an MA-node between the top two I-nodes, and a support relation connecting the bottom two I-nodes. For the purposes of identifying inference relations, we can consider either of the top two nodes acting as a premise for the conclusion at the bottom. In this case, we conflate the top two (rephrased) I-nodes, and connections to/from either are considered as being to/from both. More specifically a structure in the original annotation is considered here as being equivalent to that shown in Figure 3.2. This means that if an inference relation is (automatically) identified between *These countries, especially China are the best ever at it* and *USA don't know what they're doing when it comes to[...]*, then this is viewed as being correct.

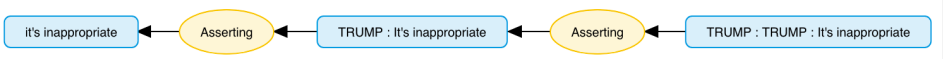


Figure 3.3: An example of reported speech in US2016G1tv.

Reported speech requires a similar simplification of the analysed structure. For example, if in the transcript we have *TRUMP: I said, "It's inappropriate."*

[20]The properties were retrieved automatically using the Argument Analytics module (Lawrence et al., 2016) of the Argument Web (Lawrence et al., 2017b) at `http://analytics.arg.tech`

then the propositional content of this is itself a locution, which has a further nested propositional content of "It's inappropriate." This can be seen in Figure 3.3. In this case, support/conflict relations can be connected to either the middle locution (if someone is, for example, questioning whether Trump *said* it is inappropriate) or, the leftmost I-node (if someone is questioning whether it is inappropriate). Here we consider the text span in the original transcript to be linked to either of these propositional contents.

While the US2016G1tv corpus will be used to test the techniques developed throughout Chapters 4-8, in Chapter 9 these techniques are combined, and the resulting combined approach evaluated against two additional widely used argumentation corpora: the Argument Annotated Essays corpus (Stab and Gurevych, 2017) described in Section 9.6.2; and, the Argumentative Microtext corpus (Peldszus and Stede, 2016)) described in Section 9.6.3.

3.3 Conclusion

One of the first challenges faced by argument mining is the lack of consistently annotated argument data. Much recent work has focused on producing annotation guidelines targeted at specific domains (e.g. Kirschner et al. (2015); Walker et al. (2014); Kiesel et al. (2015)), and whilst this has shown that data from these fields can be consistently annotated, the use of specific annotation schemes aimed at individual areas means that any techniques developed using this data are limited to that domain. The volume of data, particularly data annotated at the most fine grained level, is still far below what would be required to apply many of the techniques previously discussed in a domain independent manner. Attempts are being made to overcome this lack of data, including the use of crowdsourced annotation (Ghosh et al., 2014; Skeppstedt et al., 2018) and automatic methods to extend the data currently annotated (Bilu et al., 2015). As these efforts combine with increasing attention to manual analysis, the volume of data available should increase rapidly. Schulz et al. (2018) also offer some solace in this regard, showing how multi task learning (training models across datasets from different domains), can improve results

in domains where limited domain specific annotated data is available.

Even in cases where there is a greater volume of data, conflicting notions of argument are often problematic. In a qualitative analysis of six different, widely used, argument datasets, Daxenberger et al. (2017) show that each dataset appears to conceptualize claims quite differently. These results clearly highlight the need for greater effort in building a *framework* in which argument mining tasks are carried out, covering all aspects from agreement on the argument theoretical concepts being identified, through to uniform presentation of results and data.

The US2016tv corpus used in this thesis is the largest corpus of analysed dialogical argumentation currently available (Visser et al., 2020a), and US2016G1tv the largest corpus of dialogical argumentation completely annotated with argumentation scheme instances (Visser et al., 2021). US2016tv is annotated using Inference Anchoring Theory (IAT) (Budzynska and Reed, 2011), distinguished by being a theory of argumentation geared towards computational linguistic methods and software implementation. To facilitate machine-readability, IAT adheres to the Argument Interchange Format (AIF) standard (Chesñevar et al., 2006), a graph-based ontology that facilitates the representation argumentative structures and offers integration with a broad range of existing AIF compliant tools (Bex et al., 2013).

Chapter 4

Discourse Indicators

4.1 Introduction

The first explainable argument mining approach which we will consider in this thesis is that of using discourse indicators to determine the argumentative connections between adjacent propositions in a piece of text. Discourse indicators are explicitly stated linguistic expressions of the relationship between statements (Webber et al., 2011), and, when present, can provide a clear indication of its argumentative structure (van Eemeren et al., 2007). For example, if we take the sentence "Britain should disarm because it would set a good example for other countries", then this can be split into two separate propositions "Britain should disarm" and "it [disarming] would set a good example for other countries". The presence of the word "because" between these two propositions clearly tells us that the second is being employed as a reason for the first.

Discourse indicators have been previously used as a component of argument mining techniques. For example, in Stab and Gurevych (2014b), indicators are used as a feature in multiclass classification of argument components, with each clause classified as a major claim, claim, premise or non-argumentative. Similar indicators are used in Wyner et al. (2012), along with domain terminology (e.g. camera names and properties) to highlight potential argumentative sections of online product reviews. In Eckle-Kohler et al. (2015) a German language corpus is annotated with arguments according to the common claim-

premise model of argumentation and the connection between these annotated connections and the presence of discourse indicators (or discourse markers as they are referred to here) is investigated. The results show that discourse markers are again important features for the discrimination of claims and premises in German as well as English language texts. However, there has been little study of how well indicators perform on their own, how frequently they occur in real-world text, and how well different individual indicators map to specific argumentative relations. In this chapter we will investigate these properties of indicators, looking at indicators from existing literature, as well as those which can be identified from annotated argument data. In doing so, we will determine how reliable indicators are and what role they can play in an argument mining system.

4.2 Indicators and Argumentative Relations

There are many different ways in which indicators can appear, and a wide range of relations which they can suggest (Knott, 1996). We limit our search here to specific terms indicating support or attack relations between a pair of propositions. Specifically, we consider those indicators which show an argumentative relation between sequential propositions of the form **A** [***indicator***] **B** (e.g. "Britain should disarm *because* it would set a good example for other countries") or [***indicator***] **A B** (e.g. "*Because* we want to set a good example for other countries, we should reduce our nuclear capability"). Furthermore, we consider the relationship between indicators and the directionality of the argumentative connections (e.g. **A *because* B** suggests a support relation from the premise **B** (single underlined) to the conclusion **A** (double underlined), whereas **A *therefore* B** suggests a support relation from **A** to **B**). For this work we do not consider the more variable form of [***indicator***] **A B** as there is no clear limit on how long before **A** the indicator must occur.

In this work, two sources of candidate discourse indicators are used: an aggregation of those found in existing literature (Groarke et al., 1997; Knott, 1996)(DI-Lit), and a domain specific list extracted from relations in a set of

Relation Type	Indicators
A $\xrightarrow{support}$ B	so, therefore, accordingly, then, thus, consequently, hence, ergo
A $\xleftarrow{support}$ B	because, since, as
A $\xrightarrow{conflict}$ B	but, however, nonetheless, nevertheless, still, yet, though, whereas
A $\xleftarrow{conflict}$ B	although, except, despite, albeit

Table 4.1: Argumentative discourse indicators from existing literature.

corpora[1] (DI-Dom). In each case, we also extend these lists by including synonyms of each word identified using Synsets (groupings of synonymous words that express the same concept) from WordNet (Miller, 1995). For example, five synonyms were found for the word "therefore" ("thence", "thus", "so", "hence", and "consequently"). The original words and all synonyms were compiled into a list with duplicates removed. The indicators from DI-Lit are shown in Table 4.1.

For DI-Dom, we consider the corpora US2016D1tv and US2016R1tv, and extract those unigrams which occur between adjacent spans of text that are connected by a support or attack relation in the argumentative analysis. Common unigrams (such as "and" and "I") which appeared in more than one type of relation were removed from the final lists. Those unigrams appearing more than once for each relation type are shown in Table 4.2.

4.3 Implementation

To determine the efficacy of discourse indicators in identifying argument structure, all pairs of sequentially adjacent ADUs in the same turn were extracted from US2016G1tv (based on the manual segmentation of ADUs from the corpus). For each pair, the text between the two ADUs was tokenized into unigrams, and these tokens were then searched for each of the indicators in DI-Lit

[1]US2016D1tv and US2016R1tv, described in Section 3.2.

A $\xleftarrow{support}$ B		A $\xrightarrow{support}$ B		A $\xleftarrow{conflict}$ B		A $\xrightarrow{conflict}$ B	
Count	**Unigram**	**Count**	**Unigram**	**Count**	**Unigram**	**Count**	**Unigram**
50	because	13	so	9	well		
5	think	9	think	2	think		
3	you	7	why	2	that		
3	said	3	to	2	of		
3	know	2	which	2	all		
3	as	2	fact				
2	yes	2	which				
2	was	2	in				
2	reason						
2	now						
2	injustice						
2	first						
2	believe						

Table 4.2: The most commonly occurring unigrams between pairs of adjacent spans linked by a support or attack relation, in US2016D1tv and US2016R1tv.

and DI-Dom.

For each occurrence of each indicator, a comparison was made to the analysed argumentative structure from the original corpus, and if the corresponding support or conflict relationship was marked, then this was considered a correct identification (true positive) for that indicator, if there was no corresponding relation in the original corpus, this was a false positive. In cases where there was a relation marked in the corpus, but the indicator was not present, this was viewed as a false negative.

4.4 Results

Table 4.3 lists the top ten performing discourse indicators, sorted by the F-Score calculated using the interpretation described above. It can be seen from this table that only a relatively small number of the indicators we are searching for are actually found in the data. It is particularly surprising that indicators which are commonly mentioned in the literature as being useful for identifying

argumentative structure rarely occur: for example, "therefore" only has one occurrence within the entire debate transcript (this does indeed occur between two inferentially linked text spans).

Of those indicators which do appear more frequently in US2016G1tv, most are providing little information. For example, whilst there were 30 instances of the indicator "so" occurring between adjacent spans, only 37.5% of these instances were between spans where a support relation exists.

The one exception here is the indicator "because". This indicator appears between spans 71 times and, of these, 87.3% are connected by a support relationship. Whilst this is a promising result, and suggests that, in those cases where "because" occurs, it can tell us with high accuracy the type of connection, we can also see that using this method on its own would leave approximately 80% of support relations (as well as all conflict relations) unidentified.

Indicator	**Number**	**Precision**	**Recall**	**F-Score**
because	71	0.873	0.212	0.342
so	32	0.375	0.041	0.074
think	44	0.205	0.031	0.054
that	28	0.179	0.017	0.031
as	8	0.375	0.010	0.020
therefore	1	1	0.003	0.007
since	0	0	0	0
consequently	0	0	0	0
thus	0	0	0	0

Table 4.3: Top ten performing discourse indicators, sorted by F-Score.

These results are supported by those of earlier work (Lawrence and Reed, 2015) carried out on the Araucaria corpus (Reed et al., 2008). Focusing on the thirteen most reliable support indicators and eleven most reliable conflict indicators, Lawrence and Reed achieved an overall precision of 0.89, but a recall of only 0.04, concluding that: "discourse indicators may provide a useful component in an argument mining approach, but, unless supplemented by other methods, are inadequate for identifying even a small percentage of the

argumentative structure".

In Chapter 5, it is shown how, despite low recall, the high precision of the indicator "because" can be used to harvest additional training data, and give us another technique for identifying inferences in those cases where the connection isn't so explicitly expressed.

Chapter 5

Premise-Conclusion Topic Models

5.1 Introduction

The intuition underlying the work presented in this chapter is that there are rich and predictable thematic and lexical regularities present in the expression of human reasoning, and that these regularities can be identified in helping to extract the structure of reasoning. For example, in debates concerning abortion, arguments are carefully marshalled on both sides, with religious themes more typically appearing on one side, and feminist philosophy themes more typically on the other. For a debate on the construction of a new road, we may expect to find environmental issues on one side and economic concerns on the other. If such generalisations are possible at a coarse scale, perhaps they are similarly possible at a more fine-grained scale.

These themes are represented in terms of both the topics discussed and the language used to express them: an anti-abortion stance is likely to not just cover feminist philosophy themes in general, but to also use specific terminology more frequently, perhaps mentioning 'choice' or 'freedom' more than views expressed on the other side of the debate. When humans hear such a debate, they understand the structure of the argument being made, not only based on the content of the argument itself, but on a broad general knowledge of the topic and the way in which such arguments are commonly presented.

The argument mining technique presented here takes the commonly occurring terms in an original text and uses these terms to gather data from the web on the same topic. This large volume of additional data can be considered contextual knowledge, and is processed to find pairs of text spans which have an inferential relationship. We then use these pairs to create premise-conclusion topic models, reflecting the ways in which one topic or phraseology is commonly used to support another.

The work in Chapter 4 has shown that the discourse indicator *because* is a very reliable predictor of argument structure. Unfortunately occurrences of this indicator are also rather rare, occurring in less than 25% of argumentative inference steps. With a high-precision/low-recall technique such as is provided by this indicator, it becomes possible to process large amounts of text to extract a dataset in which we can have high confidence. This dataset can be used to capture topical regularities in the argument structure which can then be exploited in analysing text which does not benefit from the presence of indicators.

The relationship between the topics being expressed in a piece of text and the argumentative structure which it contains have been previously explored in Lawrence et al. (2014), where a Latent Dirichlet Allocation (LDA) topic model is used to determine the topical similarity of consecutive propositions in a piece of text. The intuition is that if a proposition is similar to its predecessor then there exists some argumentative link between them, whereas if there is low similarity between a proposition and its predecessor, the author is going back to address a previously made point and, in this case, the proposition is compared to all those preceding it to determine whether they should be connected. Using this method a precision of 0.72, and recall of 0.77 are recorded when comparing the resulting structure to a manual analysis, however it should be noted that what is being identified here is merely that an inference relationship exists between two propositions, and no indication is given as to the direction of this inference. This further challenging issue is addressed in Chapter 7.

5.2 Implementation

An overview of the methodology used can be seen in Figure 5.1. Starting with the manually segmented ADUs from the US2016G1tv corpus, text from these ADUs is examined in order to find the unigrams and bigrams which occur most frequently throughout the text, giving an indication of the overall theme of the text with which we are working.

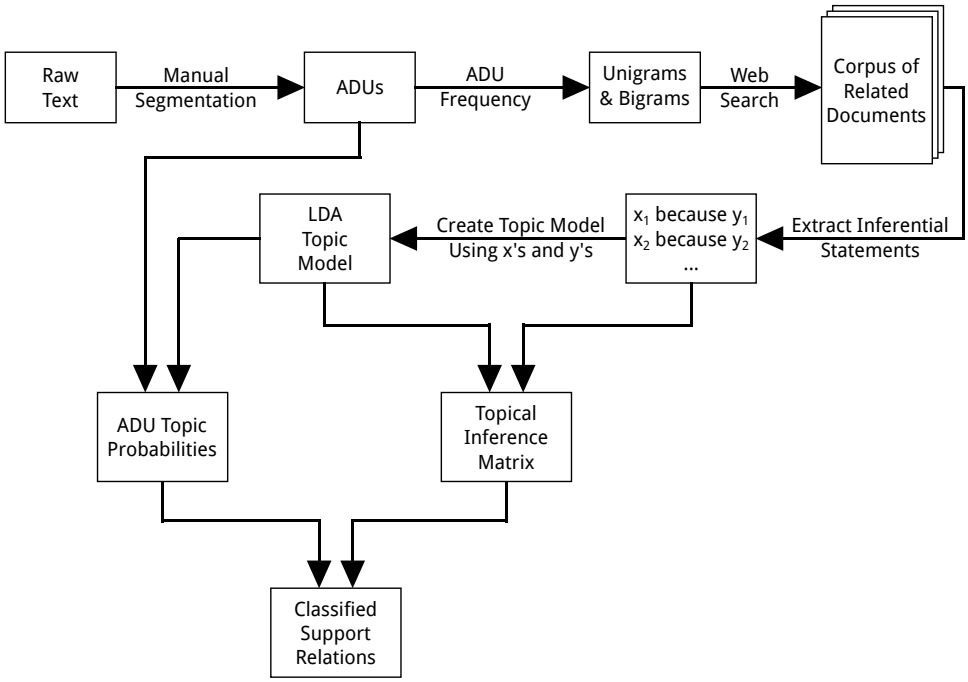


Figure 5.1: Overview of the implementation methodology for creating extended corpus, creating a topical inference matrix and classifying support relations

The next step is then to build a corpus of related documents by searching the web for those unigram and bigram terms identified as being indicative of the theme. From this extended corpus, we then extract sentences which contain an inferential relationship by searching for those discourse indicators which we have found to have the highest precision. This search results in a large collection of pairs of text fragments where one element of the pair is a premise supporting the other, that is a conclusion.

Using these fragments as documents, we then generate a Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003), and from this create a matrix capturing the probability of support between each of the identified topics. By matching pairs of ADUs from the original text against the probabilities in

this matrix, we are then able to determine the probability that there is an inferential relationship between them, and by thresholding these values, we can then categorise ADU pairs as being 'inferential' or 'non-inferential'.

An alternative approach would be to use the premise/conclusion dataset as training data for a supervised machine learning approach. This is limited by the fact that we only obtain positive examples, and, whilst techniques such as PU-learning (Learning from Positive and Unlabelled examples) (Liu et al., 2003) provide a way of dealing with only positively labelled data, we do not have sufficient quantities of unlabelled examples for these techniques to be applied. In future work, the ability to identify arbitrary ADUs in text could be used to extract large volumes of unlabelled examples, and such approaches may then become more suitable.

5.2.1 Obtaining Premise/Conclusion Pairs

The first step in the pipeline described above is to determine the overall theme of the text being analysed. This is achieved using an unsupervised keyword extraction approach. Firoozeh et al. (2020) provide a comprehensive overview of keyword extraction techniques comparing a range of both supervised and unsupervised methods. For this work, we require an unsupervised method as there is no annotated keyword data available. In particular we utilise a statistical approach leveraging the existing segmentation of the US2016G1tv corpus into ADUs, and calculate the number of unique ADUs in which each unigram or bigram appeared. This list is then sorted and filtered to remove common stop words. The top ten from each resulting list of terms can be seen in Table 5.1 and Table 5.2.

Having identified keywords describing the topic, a corpus of related documents was created by searching the web for combinations of these terms. The top twenty terms of each kind were combined into search queries by taking all possible combinations of two and three unigrams as well as each bigram both on its own and paired with each unigram. Using these queries, the first 100 Google search results for each were compiled. After filtering the list of related documents to remove duplicates, a total of 8,684 pages remained.

Unigram	**Count**
Trump	372
Clinton	302
people	83
USA	77
country	64
tax	53
jobs	48
ISIS	44
Obama	42
nuclear	36

Table 5.1: Top ten unigrams by number of ADUs in which they appear

Bigram	**Count**
trade deals	15
New York	14
Trans Pacific	11
tax returns	11
united states	11
african american	10
new jobs	10
nuclear weapons	10
cyber attacks	10
middle east	9

Table 5.2: Top ten bigrams by number of ADUs in which they appear

Source	Documents	Inferential Sentences
Web search	8,684	5,897
News Articles	1,012	803
Speeches	118	1,517
Twitter	6,000	103
Total	**15,814**	**8,410**

Table 5.3: Number of documents and inferential sentences for each data source

Although the pages identified in the previous step are high ranking search results for the terms identified, such pages commonly contain material unrelated to the topic, for example, advertisements and summaries of other articles. In order to extract those sections of the documents most likely to contain the body of an article, the Python *Beautiful Soup* library[1] was used to parse the HTML and extract those sections comprising consecutive paragraphs of text.

These web search results were further supplemented by three additional sources: 1,012 news articles covering the political candidates during the year leading up to US Presidential Election extracted from the PolNeAR corpus (Newell et al., 2018); 36 speeches made by Hillary Clinton and 82 speeches made by Donald Trump, taken from the Clinton-Trump corpus (Brown, 2017); and, 3,000 Twitter posts made by the official accounts of each candidate ("realDonaldTrump" and "HillaryClinton") in the run up to the election.

The texts from all sources were split into sentences, using the NLTK[2] tokeniser, and each of the resulting sentences searched for the presence of a discourse indicator. Previous work, including (Lawrence and Reed, 2015) and that described in Chapter 4, has shown "because" to be by far the most reliable indicator of inference relations, and, as such, we limit the search here to sentences containing this word. Completing this search gives a total of 8,410 inferential sentences of the form *conclusion* ***because*** *premise*. The number of such sentences extracted from each source is shown in Table 5.3.

Though the indicator "because" does not offer 100% precision, including

[1] `http://www.crummy.com/software/BeautifulSoup/`

[2] `http://www.nltk.org/`

some non-argumentative examples, the noise is mitigated by the way in which the resulting pairs are subsequently used. The use of the topic models described in the next section means that we neither need *all* of the inferential relations contained within our search results, nor for *every* premise-conclusion pair to be correctly labelled as such. The models which we produce may have a small amount of noise generated by false-positives, but these either comprise topics which are not then matched to elements from the original text, or add a small number of lower importance terms to a valid topic.

5.2.2 Creating the Topical Inference Matrix

To extract the topical nature of the premise-conclusion pairs previously identified, a Latent Dirichlet allocation (LDA) topic model was created using the Python gensim library[3]. To produce this topic model, the sentences were first split where the indicator occurred, giving two documents from each sentence (one representing a premise, and the other, the conclusion). To determine the optimal number of topics, we experimented with values in the range 2–45, calculating the topical coherence (Newman et al., 2010) for each (see Figure 5.2). Based on these results, using twenty topics appears to provide high coherence whilst avoiding the risk of repetition by selecting a higher number. As such, we generated our final model with twenty topics and using fifty passes over the supplied corpus.

From the probability distributions for each pair of conclusion (C) and premise (P), a topical inference matrix (T) was created, where the i,jth entry in the matrix corresponds to the product of probabilities that the premise has topic i and the conclusion topic j. For example, in the simplest case, if there is a probability of 1.0 that the premise has topic m and the conclusion topic n, then the matrix will contain 1.0 at m,n and zero for all other possible pairings. So, given topic distributions θ^C for the conclusion, and θ^P for the premise, T is defined thus:

$$t_{i,j} = \theta_i^P \theta_j^C \tag{5.1}$$

[3] `https://radimrehurek.com/gensim/`

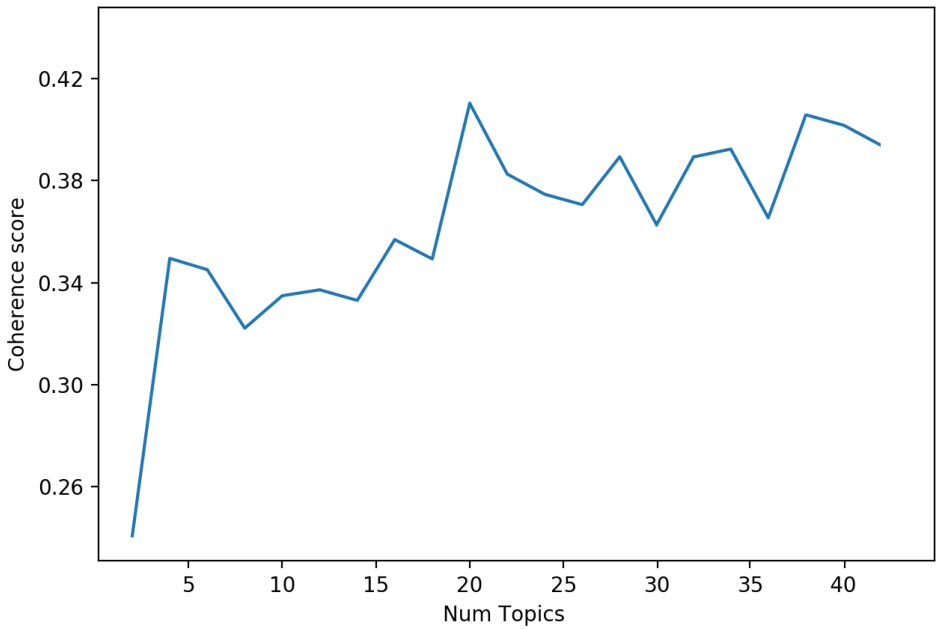


Figure 5.2: Topical coherence for a range of different numbers of topics

The individual matrices for each premise/conclusion pair can then be summed together, and the resulting values normalised by dividing by the maximum value in the matrix to give values between 0 and 1. The result is an overall topical inference matrix (U) where each value represents the likelihood that a premise matching topic i supports a conclusion matching topic j. A heatmap representation of this matrix can be seen in Figure 5.3. This representation highlights some interesting points about the relationship between premise/conclusion topics and argument structure. Firstly, there is a clear pattern along the diagonal top left to bottom right (i.e. premise/conclusion pairs where both have the same topic). This tendency for premises and conclusions to be topically similar is explored further in Chapter 6. Secondly, the matrix is not symmetrical between premises and conclusions. For example, there are a substantial number of conclusions that correspond to topics 16 and 17, whilst these are much less common for premises. This lack of symmetry in topical distribution between premises and conclusions suggests that, given a pair of ADUs, the matrix may be able to help us determine not only whether they are inferentially connected, but also which is premise and which conclusion. Both of these tasks are investigated in the following section.

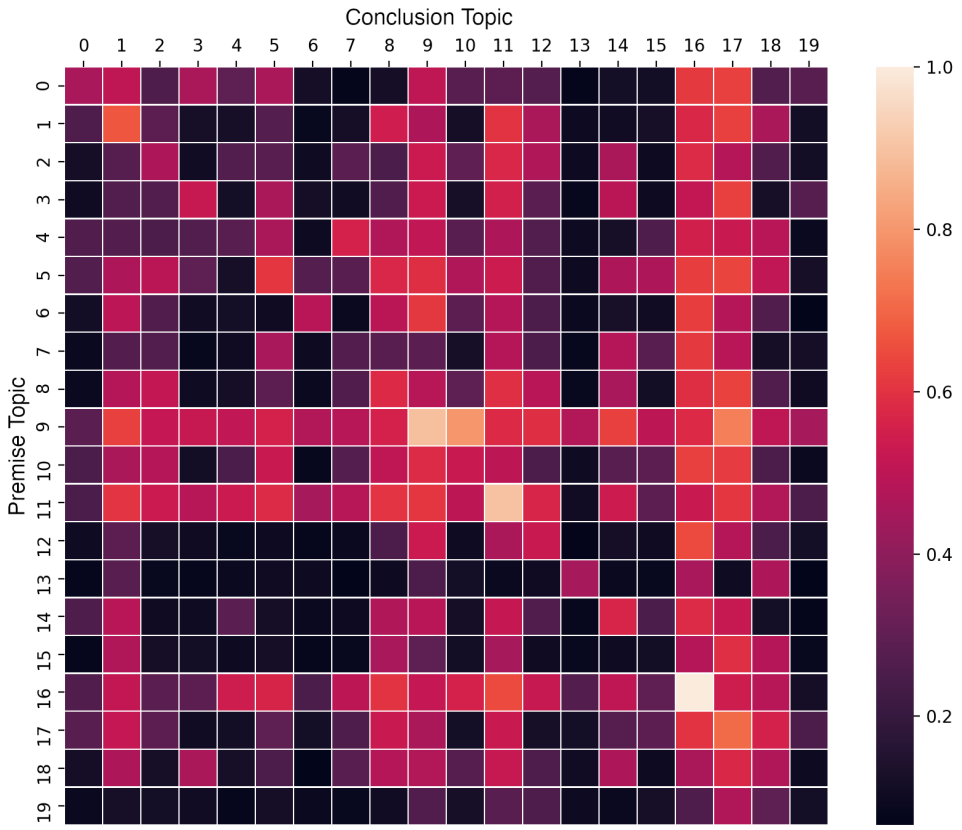


Figure 5.3: Heatmap of the topical inference matrix for the 2016 US Presidential Election

5.3 Experiments

In order to test our original hypotheses that the thematic regularities present in the expression of human reasoning can be identified and used to help determine the structure of that reasoning, a number of experiments were carried out to explore the effectiveness of using this data to determine both the direction of inference between two ADUs that are known to have an inferential relationship, and the connectedness of pairs of arbitrary ADUs.

5.3.1 Using the Topical Inference Matrix to determine directionality

Our test data is comprised of the 505 premise-conclusion pairs from the US2016G1tv corpus. As an initial experiment, we investigated how well the produced topical inference matrix could determine the direction of the inference between

these pairs. This was achieved by creating a test set containing each pair (a,b) and its reverse (b,a).

Two alternative methods were tested to classify these pairs as being 'inferential' or 'non-inferential'. In each case, the topic probabilities for the ADUs were first inferred from the LDA model and a score determined as to whether there was an inferential relationship. For the first method, MaxTopic, the score was calculated by taking the highest probability topic for each ADU and using these to look up the corresponding value in the overall topical inference matrix (U):

$$S_{MaxTopic}(P,C) = u_{\arg\max_i(\theta_i^P),\arg\max_j(\theta_j^C)} \tag{5.2}$$

For the second method, TopicDist, the values in the overall topical inference matrix (U) were multiplied by the corresponding probabilities for each item in the pair and then summed to give an overall score.

$$S_{TopicDist}(P,C) = \sum_{i=1}^{n}\sum_{j=1}^{n} u_{i,j}\theta_i^P\theta_j^C \tag{5.3}$$

For each of these two methods, the resulting scores were then compared against the mean of all values in the matrix (mean = 0.46), over which a pair would be classified as being 'inferential', and below which, 'non-inferential'. It is not possible to choose a fixed threshold value empirically based on results from a development data set as the variation in probabilities in different matrices is unknown. Experiments were carried out using the US2016D1tv and US2016R1tv corpora to compare different calculated threshold values (median, mean, and the kth largest value in the matrix for all possible values of k), and the mean was found to give the maximum f-score of all values tried across both corpora.

The results for directionality can be seen in Table 5.4. The results show an improvement over the random baseline for both methods, however the improvement in precision is low when just looking at the highest scoring topic. One reason for this is that a reasonable percentage of pairs (107 out of 505) have the same highest scoring topic for both items (i.e. a conclusion is being supported by a premise that is closely related). When these same topic

Method	**Precision**	**Recall**	F_1**-score**
Random Baseline	0.5	0.5	0.5
MaxTopic	0.53	0.86	0.66
TopicDist	**0.63**	**0.86**	**0.73**

Table 5.4: Results for the MaxTopic and TopicDist methods to determine directionality of inferential connections compared to the random baseline

pairs are removed, the precision increases to 0.60, comparable to the results for the weighted topic distribution. The results for using the weighted topic distribution are better, and suggest that even in cases where the main topic is similar, there is enough of a difference in the secondary topics to determine the directionality of the pair.

5.3.2 Using the Topical Inference Matrix to determine connectedness

The second experiment performed looked at whether the produced topical inference matrix could determine inferential connections between arbitrary pairs of ADUs. For this task, a dataset was created containing the known 505 premise-conclusion pairs from US2016G1tv and an equal number of unconnected ADU pairs randomly selected from US2016G1tv. The same two methods of classifying these pairs as being 'inferential' or 'non-inferential' were used as in the first experiment, and the results can be seen in Table 5.5.

The results show that the precision is increased for classifying pairs as being connected over the previous results for directionality.

Method	**Precision**	**Recall**	F_1**-score**
Random Baseline	0.5	0.5	0.5
MaxTopic	0.68	0.91	0.77
TopicDist	**0.69**	**0.90**	**0.78**

Table 5.5: Results for the MaxTopic and TopicDist methods to determine connectedness of ADU pairs

5.4 Discussion

The results we have presented show in all cases that there is some correlation identified between the topics that a pair of ADUs have, and the nature of their potential inferential relationship. By looking at the topics of each item in the pair, we have been able to determine both connectivity and directionality of inference. Overall, the results are better for identifying connectedness than directionality, predominantly resulting from higher similarity in topics for which the ADUs are connected (in a significant percentage of cases the maximum probability topic was the same).

Currently, the identification of relationships is limited to inferential relationships, and one area of development would be to extend this by examining those discourse indicators which show a conflict relationship. Additionally, no account is taken of the polarity or sentiment of the ADUs. Where we have a conclusion, 'C', and a premise, 'P', then there would be a high topical similarity between P and 'not P', and as such, an inference relationship would be assigned between them. This problem could be overcome by applying negation detection techniques (Jia et al., 2009) to the ADUs as a preliminary step, and where there is negation of one item in the pair, replacing an inference relationship with conflict.

Although we focus on identifying patterns of inference within a single debate, there is nothing intrinsic to the approach that makes it a better fit for this domain than any other. The use of the indicator "because" has been shown to give high precision across multiple domains and corpora (Lawrence and Reed, 2015, 2017a). The automatic determination of the domain being discussed requires only the original text, and from this we are able to build a dataset specific to that domain which, due to the reliability of discourse indicators, contains domain specific pairs that we can say with high confidence have an inferential relationship.

5.5 Conclusion

The work in this chapter has demonstrated how by automatically creating large, high-confidence datasets of inferential pairs related to a specific topic, we can closely mirror one of the ways in which humans understand the complex interactions between the individual propositions expressed in a debate.

The approach presented is effective in tackling the challenging high-level pragmatic task of identifying both connectedness and directionality between argumentative discourse units. This outcome represents strong performance for this level of task (cf., for example, (Feng and Hirst, 2011; Peldszus, 2014)), giving results comparable to those of (Palau and Moens, 2009), where each argument sentence was classified as either premise or conclusion with F_1-scores of 0.68 for classification as premise and 0.74 for conclusion. Furthermore, where existing approaches are often constrained in their generality by a lack of appropriately annotated, domain-specific, data, the same requirement does not apply in this case.

The results show a clear link between the words used to express an argument and its underlying structure, and strongly support the intuition that understanding the structure of an argument requires not only consideration of the text itself, but contextual knowledge and understanding of the broader issues.

Chapter 6

Similarity and Topical Changes

6.1 Introduction

In this chapter, we consider how various measures of the similarity between propositions map to their argumentative relationship. It seems intuitive that a premise and its associated conclusion may often share a large number of words in common, as in the following example from US2016G1tv:

> **Premise:** they lost plenty of money on investing in a solar company
> **Conclusion:** that was a disaster to invest in a solar company

or, be semantically similar without sharing very many common words, for example:

> **Premise:** We also have to make the economy fairer
> **Conclusion:** CLINTON also wants to see more companies do profit-sharing

Indeed, previous work (Lawrence et al., 2014; Boltužić and Šnajder, 2014; Wachsmuth et al., 2018) has shown that various similarity measures can be used to successfully determine not just individual argumentative relationships, but the entire argumentative structure contained within a text. We will look at this approach in more detail in Section 6.5, but first, we consider a range of similarity measures in isolation, measuring their efficacy in determining both long and short distance inferential relationships.

6.2 Similarity Measures

6.2.1 Lexical Similarity Measures

ROUGE-1, ROUGE-2 and ROUGE-L: ROUGE, or Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004), is a set of metrics designed for evaluating automatic summarisation by comparing an automatically produced summary against a reference summary. The ROUGE-N metric compares an automatic summary with a reference summary using the n-gram overlap between the two documents. If 1-grams (individual word tokens) are used to compare the documents, then the metric is called ROUGE-1, for 2-grams (pairs of consecutive word tokens) it is called ROUGE-2 and so on. More formally, ROUGE-N is the n-gram recall between a candidate summary and a set of one or more reference summaries, and is calculated as follows:

$$ROUGE\text{-}N = \frac{\sum_{S \in S_H} \sum_{g_n \in S} Count_{match}(g_n)}{\sum_{S \in S_H} \sum_{g_n \in S} Count(g_n)} \tag{6.1}$$

where: S_H is the set of manual summaries; g_n is an n-gram; $Count_{match}(g_n)$ the maximum number of n-grams co-occurring in the candidate summary and set of reference summaries; and, $Count(g_n)$ the number of n-grams occurring in the candidate summary.

ROUGE-L measures the longest matching sequence of words using Longest Common Subsequence (LCS) (Lin and Och, 2004). An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, a predefined n-gram length is not required. ROUGE-L uses the F-measure to estimate the similarity between a reference summary, X, of length m and a candidate summary, Y, of length n, calculated as follows:

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{6.2}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{6.3}$$

$$ROUGE\text{-}L = F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{6.4}$$

where: $LCS(X, Y)$ is the length of the longest common subsequence of X and Y; and $\beta = P_{lcs}/R_{lcs}$.

Levenshtein Edit Distance: The Levenshtein edit distance (Levenshtein, 1966) is a metric for measuring the difference between two sequences in terms of the minimum number of single-element edits (insertions, deletions or substitutions) required to change one sequence into the other.

The Levenshtein distance between two strings a and b (of length $|a|$ and $|b|$ respectively) is given by $lev_{a,b}(|a|, |b|)$ where:

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & \text{if } min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + k \end{cases} & \text{otherwise.} \end{cases} \tag{6.5}$$

and, where $k = 0$ if $(a_i = b_j)$, or 1 otherwise. Here the first element in the minimum corresponds to deletion (from a), the second to insertion and the third to match or mismatch, depending on whether the respective symbols are the same.

As described in the equation above the Levenshtein edit distance is looking at the distance between strings in terms of character operations, in this form it is most commonly used for spelling correction (find the closest word from a given vocabulary), however, by changing the type of sequence and constituent elements, this metric has found applications ranging from DNA sequencing (Buschmann and Bystrykh, 2013) to plagiarism detection (Gipp and Beel, 2010).

In this case, however, the distance is calculated by considering a and b as the two given propositions and the elements under consideration being the stemmed words that they contain, rather than letters. To obtain a value in the range 0-1, the Levenshtein distance is divided by the maximum possible

distance between the strings, that is, divided by max(length(a), length(b)). Using the Levenshtein edit distance in this way allows for situations where word ordering is substantially altered, but the meaning is still similar.

6.2.2 Semantic Similarity Measures

WordNet: To calculate semantic similarity using WordNet (Miller, 1995), an algorithm based on the general template method for semantic similarity given in Mihalcea et al. (2006) was used. Mihalcea et al. propose combining metrics of word-to-word similarity and word specificity into a formula that is a potentially good indicator of the semantic similarity of two input texts. In this case, this is achieved using WordNet path similarity[1] as the word-to-word similarity metric, and Inverse Document Frequency (IDF) (Sparck-Jones, 1972) as the word specificity metric. With these metrics similarity can then be calculated using the equation below:

$$sim(T_1, T_2) = \frac{1}{2}(\frac{\sum_{w \in T_1}(maxSim(w, T_2) * idf(w))}{\sum_{w \in T_1} idf(w)} + \frac{\sum_{w \in T_2}(maxSim(w, T_1) * idf(w))}{\sum_{w \in T_2} idf(w)}) \tag{6.6}$$

That is, for each word w in segment T_1 we identify the word in segment T_2 that has the highest path similarity score ($maxSim(w, T_2)$), this is then weighted with the corresponding word specificity, summed up, and normalised with the length of each text segment. This process is repeated with each word in T_2, and the resulting similarity scores are combined using a simple average. This similarity score has a value between 0 and 1, with a score of 1 indicating identical text segments, and a score of 0 indicating no semantic overlap between the two segments.

Word Vectors: The final two semantic similarity approaches tested were

[1]Path similarity is available as part of the NLTK WordNet interface (`http://www.nltk.org/howto/wordnet.html`), and is inversely proportional to the number of nodes along the shortest path between the synsets, with the maximum value being 1 when the two synsets are the same, and the minimum being 0.

implemented using pre-trained models. Word vectors were explored using word2vec (Mikolov et al., 2013), an efficient neural approach to learning high-quality embeddings for words. Specifically, the pre-trained skip-gram vectors trained on a Google News dataset[2] were used. This model contains 300-dimensional vectors for 3 million words and phrases.

To determine similarity between propositions, the centroid of the word embeddings was located by averaging the word2vec vectors for the individual words in the proposition. The cosine similarity between centroids was calculated to represent the proposition similarity.

Document Vectors: Document vectors were implemented using a doc2vec (Le and Mikolov, 2014) distributed bag of words(*dbow*) model to represent every proposition as a vector with 300 dimensions. Again, the cosine similarity between vectors was then calculated to represent the proposition similarity.

6.2.3 Topical Similarity Measures

Latent Dirichlet Allocation: As in Lawrence et al. (2014), a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) is also used here to determine similarity. Later work (Lawrence and Reed, 2015) has suggested that LDA performs less successfully on shorter text spans, though we include this model here both for completeness, and to investigate whether this is indeed the case for the US2016G1tv corpus.

LDA is a generative model which conforms to a Bayesian inference about the distributions of words in the documents being modelled. Each topic in the model is a probability distribution across a set of words from the documents. Once the model is generated, a specific individual document can be compared to it, in order to obtain scores for how well the document matches each identified topic. In order to obtain a similarity score for two documents, we can compute the topic scores for each, and then calculate the Euclidean distance between these scores.

[2] `https://code.google.com/archive/p/word2vec/`

6.3 Similarity Experiments

6.3.1 Similarity and Argumentative Relations

The first experiment performed using the similarity measures described in the previous section investigates the link between similarity and argumentative relations. In order to do this, the similarity scores for all pairs of connected propositions in the US2016G1tv corpus were calculated, and the average of these compared to the average similarity score for all non-connected adjacent propositions. In this case, we consider any two propositions connected by either a support (RA) or conflict (CA) relation as being related. There is no way, from similarity alone, to tell which of these two relations holds in each case. The results of this comparison can be seen in Table 6.1.

Technique	**Un-Related**	**Related**	**Significance**
Lexical Similarity Measures			
ROUGE-1	0.0662	0.1486	$p < 0.001$
ROUGE-2	0.0028	0.0377	$p < 0.001$
ROUGE-L	0.0515	0.1190	$p < 0.001$
Levenshtein	0.1214	0.2238	$p < 0.001$
Semantic Similarity Measures			
WordNet	0.1625	0.1711	$p > 0.05$
Word Vectors	0.2111	0.3297	$p < 0.001$
Document Vectors	0.1114	0.1362	$p < 0.05$
Topical Similarity Measures			
LDA	0.4469	0.4638	$p > 0.05$

Table 6.1: Average similarity scores for related and un-related propositions with significance of difference calculated using Student's t-test.

The first thing that can be observed from these results is that there is a highly significant ($p < 0.001$) difference between the similarities of related and un-related propositions, as calculated by the majority of measures. The exceptions to this are the results obtained using Document Vectors (which

were, nonetheless, still significant, $p < 0.05$), and those for WordNet and LDA, which showed no significant difference between the average values. In the case of WordNet, this seems most likely to be caused by the IDF weighting - it is possible that there are words which occur in a relatively high number of propositions, but where these are all related, equally that less common words may not be a useful indicator. For LDA, this result seems most likely due to the length of documents, as discussed in the previous section. As WordNet and LDA performed significantly more poorly, they were excluded from subsequent experiments.

The ROUGE metrics generally have lower scores than the rest, with ROUGE-2 being notably lower than the other two. The lower average scores for ROUGE-2 are perhaps to be expected, repeated bigrams are indeed likely to be less common than repeated unigrams (though despite these lower scores, ROUGE-2 shows the "most significant" difference of the three). And word repetition is perhaps less likely than similar semantic meaning, though still a useful indicator when it does occur. Word Vectors also outperform Document Vectors, suggesting that these are better capturing the meaning of the propositions. However, as data from which they are produced is different in each case, it is unclear whether this is a result of the technique or of the data.

6.3.2 Similarity and Adjacency

It could be suggested that one cause for the higher similarity between related propositions is that they often occur close together in a dialogue, and, as such, are more likely to be similar due to their proximity rather than because of any connection to argument structure. It is indeed the case that ∼20% of related propositions are sequentially adjacent.

In order to investigate this further, the same similarity measures were used to calculate average scores again, but looking this time at related and unrelated propositions that are sequentially adjacent. The results of these calculations are shown in Table 6.2.

Whilst the averages are higher in all cases for adjacent propositions, it can be seen from the results table that there is still a significant difference

Technique	Un-Related	Related	Significance
Lexical Similarity Measures			
ROUGE-1	0.1911	0.3336	$p < 0.05$
ROUGE-2	0.0142	0.0393	$p < 0.001$
ROUGE-L	0.1708	0.2771	$p < 0.001$
Levenshtein	0.2214	0.3823	$p < 0.001$
Semantic Similarity Measures			
Word Vectors	0.3059	0.5999	$p < 0.001$
Document Vectors	0.2341	0.3062	$p < 0.05$

Table 6.2: Average similarity scores for **adjacent** related and un-related propositions.

(though slightly reduced in the case of ROUGE-1) between related and un-related propositions for all of the techniques. From this we can conclude that even for sequential propositions, those that have an argumentative relation connecting them are generally more similar than those that don't.

6.4 Long distance

So far the results have shown that there is a significant difference in similarity between related and un-related propositions, and the scale of difference remains similar when looking at just those propositions that are adjacent. But, what about long range relations? In this final experiment before moving on to apply these results as an argument mining technique, we look at whether the similarity is significantly different between propositions that have a relation to each other and are situated further apart in the dialogue (with > 5 propositions between them[3]), and those similarly situated, but with no argumentative relation. The results of these calculations are shown in Table 6.3.

The most notable results from the table are those for the semantic measures, which, now looking at distant relations, are showing a great increase

[3]The distance of 5 was selected as this value ensures that, at a minimum, all pairs are in different dialogue turns.

Technique	**Un-Related**	**Related**	**Significance**
Lexical Similarity Measures			
ROUGE-1	0.0655	0.0961	$p < 0.05$
ROUGE-2	0.0028	0.0057	$p < 0.001$
ROUGE-L	0.0499	0.0810	$p < 0.05$
Levenshtein	0.1212	0.1921	$p < 0.05$
Semantic Similarity Measures			
Word Vectors	0.2276	0.5291	$p < 0.001$
Document Vectors	0.1114	0.4332	$p < 0.001$

Table 6.3: Average similarity scores for **distant** (> 5 propositions apart) related and un-related propositions.

in significance for the difference between related and un-related propositions. Whilst this result may seem surprising at first glance, it reflects the likelihood that when a speaker is referring back to a previous point in the dialogue they may paraphrase the original point retaining its meaning (resulting in higher semantic similarity), but using different words to when it was originally uttered (resulting in possibly lower lexical similarity). Misra et al. (2016) look at this aspect of using various similarity measures to identify *argument facets*, or groups of paraphrased arguments, with the aim of clustering and grouping similar arguments and producing argument facet summaries as a final output.

6.5 Argument mining with similarity

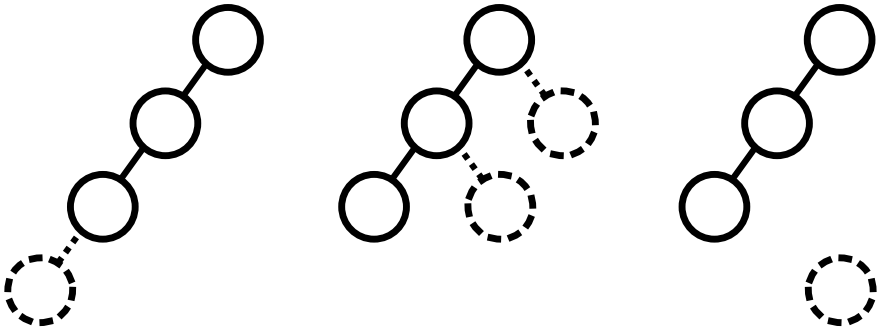
Having investigated the relationship between various measures of semantic and lexical similarity, we now move on to apply these to determining argument relations in text. To do this, we adapt the "Topical Similarity" argument mining technique presented in Lawrence et al. (2014). This technique relies on two assumptions: firstly that the argument structure to be determined can be represented as a tree, and secondly, that this tree is generated depth first. That is, the conclusion is given first and then a line of reasoning is followed supporting this conclusion. Once that line of reasoning is exhausted,

the argument moves back up the tree to connect to one of the previously made points. If the current point is not similar to any of those made previously, then it is assumed to be un-connected, and possibly the start of a new topic.

Although the assumption of a tree structure does not hold for all arguments, it is the case for around 95% of the argument analyses contained in AIFdb, and 80% of arguments in the Consumer Debt Collection Practices (CDCP) corpus as reported by Niculae et al. (2017). Similarly, not all arguments are presented in a depth first manner, though this is indeed the most common ordering. For example, Stab and Gurevych (2017) use a heuristic baseline classifying the first argument component in each body paragraph as a claim, and all subsequent components in the paragraph as premises for this claim. This baseline is shown to give an F1-score of 0.74 on the Argument Annotated Essays Corpus (AAEC) (Stab and Gurevych, 2014a).

Based on these assumptions the argumentative structure is determined by looking at how similar each proposition is to its predecessor. If they are sufficiently similar, it is assumed that they are connected and that the line of reasoning is being followed. If they are not sufficiently similar, then it is first considered whether we are moving back up the tree, and the current proposition is compared to all of those statements made previously and connected to the most similar previous point. Finally, if the current point is not sufficiently similar to any of those made previously, then it is assumed to be disconnected from the existing structure. This process is illustrated in Figure 6.1.

The question that arises from this description is, what is meant by "sufficiently similar" when considering these possible connections. To get a feel for the answer to this, it is necessary to look at the tables of average similarities given in the previous section. For example, with the ROUGE-1 technique, related adjacent propositions have an average score of 0.3336, and un-related adjacent propositions a score of 0.1911. It seems that to decide if a proposition is related to its predecessor, setting a threshold somewhere between these values is a good place to start. Clearly the exact value of the thresholds can be changed to prioritise precision or recall as required by the task at hand, something which will become useful in Chapter 9 when different argument



Step 1: The similarity of a new proposition to its immediate predecessor is calculated. If the new proposition is sufficiently similar, this is viewed as a continuation of the previous line of reasoning and the two are connected.

Step 2: If the new proposition is not sufficiently similar to its immediate predecessor, the similarity to all previous propositions is calculated. The most similar previous proposition is then selected and, if it is sufficiently similar to the new proposition, a connection is made.

Step 3: If the new proposition is not sufficiently similar to any of the previous propositions, it is viewed as the start of a new line of reasoning, disconnected to the existing argument structure.

Figure 6.1: The steps involved in determining how the argument structure is connected using the "Topical Similarity" argument mining technique presented in Lawrence et al. (2014). The dashed lines represent potential connections for each step.

mining techniques are used in combination. For the purposes of comparing the different techniques here, however, the threshold was selected to give the best balance between precision and recall and was calculated as follows:

$$Threshold = Avg_{\neg rel} + (Avg_{rel} - Avg_{\neg rel})\frac{stdev_{rel}}{stdev_{rel} + stdev_{\neg rel}} \tag{6.7}$$

where Avg_{rel} and $Avg_{\neg rel}$ are the averages for related and non-related propositions respectively, and $stdev_{rel}$ and $stdev_{\neg rel}$ are the standard deviations for each. In effect this gives a point between the two averages weighted according to the standard deviation.

This same threshold is applied for both adjacent connections and long distance connections using the figures in Tables 6.2 and 6.3. For example, using ROUGE-1 the threshold is 0.204 for adjacent propositions, and 0.071 for long distance.

The results obtained from applying this approach with these threshold values calculated for each similarity technique are shown in Table 6.4. The last row of this table, 'Any', gives the results obtained when any of the similarity techniques listed above are over their respective threshold. This 'Any' measure

Technique	**Precision**	**Recall**	**F1-Score**
Lexical Similarity Measures			
ROUGE-1	0.67	0.59	0.63
ROUGE-2	0.75	0.67	0.71
ROUGE-L	0.6	0.64	0.62
Levenshtein	0.62	0.68	0.65
Semantic Similarity Measures			
Word Vectors	0.73	0.63	0.68
Document Vectors	0.71	0.6	0.65
Combined			
Any	**0.81**	**0.74**	**0.77**

Table 6.4: Precision, recall and F1-Score for identifying argumentative relations using a range of similarity techniques.

captures the different ways in which argumentatively related propositions can be similar to each other. For example, connected propositions may have similar meaning but use different words, or may have different meaning, but be related to the same topic. The table shows that the results in this last row are the best of all approaches, this is a situation perhaps hinted at, and supporting the intuition stated in, the introduction to this chapter where it is suggested that argumentatively related propositions may share a large number of words in common, or be semantically similar without sharing very many common words.

6.6 Conclusion

In this chapter a range of similarity measures have been considered, these include: lexical measures capturing the scenario where a premise and its associated conclusion share a large number of words in common; semantic measures capturing the scenario where the words used are different, but the meaning being conveyed is similar; and, topical measures where the exact words and their meaning is not shared between premise and conclusion, but they are still both

talking about the same topic.

The results show that there is a highly significant ($p < 0.001$) difference between the similarities of related and un-related propositions, as calculated by the majority of measures. We have also seen that this similarity is not just due to the proximity of premise and conclusion meaning that they are more likely to be similar.

Finally we have seen that similarity measures can be used to to determine argument relations in text. Whilst the F1-score of 0.77 is promising for this task, these predicted relations are un-typed and un-directed. It will require further combination with other argument mining techniques for these results to achieve their full potential, as will be seen in Chapter 9.

Chapter 7

Graph Properties

7.1 Introduction

In this chapter, we consider the insights that can be gained by considering large scale argument networks as a whole. We present two metrics, *Centrality*, which can be viewed as how important an issue is to the argument as a whole (how many other issues are connected to it), and *Divisiveness*, how much an issue splits opinion (how many other issues are in conflict with it and the amount of support which the two sides have).

We first show how these metrics can be calculated from an annotated argument structure and then show how they can be automatically approximated from the original text. We can then use this automatic approximation to determine the argumentative structure of un-annotated text, by using the centrality and divisiveness scores for each text span to help decide how they should be connected. In Section 7.4, we combine this approach with existing argument mining techniques and show how the identification of properties of argumentative relations can be improved by considering the larger context in which these relations occur.

Despite the rich heritage of philosophical research in argumentation theory (van Eemeren et al., 2014; Chesñevar et al., 2006), the majority of argument mining techniques explored to date have focused on identifying specific facets of the argumentative structure rather than considering the complex network of interactions which occur in real-life debate. For example, existing approaches

have considered: classifying sentences as argumentative or non-argumentative (Moens et al., 2007); classifying text spans as premises or conclusions (Palau and Moens, 2009); classifying the relations between specific sets of premises and their conclusion (Feng and Hirst, 2011); or classifying the different types of premise that can support a given conclusion (Park and Cardie, 2014).

The approach presented in this chapter considers large scale argument networks as a whole, looking at properties of argumentative text spans that are related to their role in the entire argumentative structure. In our automatic determination of *Centrality* and *Divisiveness*, we first construct a graph of similarity between text spans and then use eigenvector centrality to determine those which are most central. For *Divisiveness*, we then look at the sentiment polarity of each text span compared to the rest of the corpus to measure how many others are in conflict with it and the amount of support which the two sides have.

7.2 Large-Scale Argument Graph Properties

The argument graphs described in Chapter 3 allow us to look at the structure of the debate as a whole rather than focusing on the properties of individual relations between propositions. Where many argument corpora consist of multiple smaller texts that have no connections between them, the argument structure of US2016G1tv covers the entire debate, with links between all related parts no matter how far apart in the debate they occur.

In this section we look at two measures, *Centrality* and *Divisiveness*, that individual propositions (I-nodes) exhibit which can only be interpreted when considering the broader context in which they occur. Whilst there are certainly other measures that could be applied to an argument graph highlighting interesting features of the arguments being made, we have selected these two metrics as they can both be calculated as properties of the argument graph and approximations can be determined directly from the original text. In Section 7.3, we describe methods to determine these approximations directly from the original text. By first calculating them directly we can then reverse the

process of determining them from the argumentative structure, cutting the manual analysis out of the loop and allowing us to determine the argumentative structure directly. In Section 7.4, we look at how this approach can be used to improve the accuracy of extracting the full argumentative structure directly from un-annotated text.

7.2.1 Centrality

Central issues are those that play a particularly important role in the argumentative structure. For example, in Figure 7.1, we can see that the node "CLINTON knows how to really work to get new jobs..." is intuitively more central to the dialogue, being the point which all of the others are responding to, than the node "CLINTON's husband signed NAFTA...".

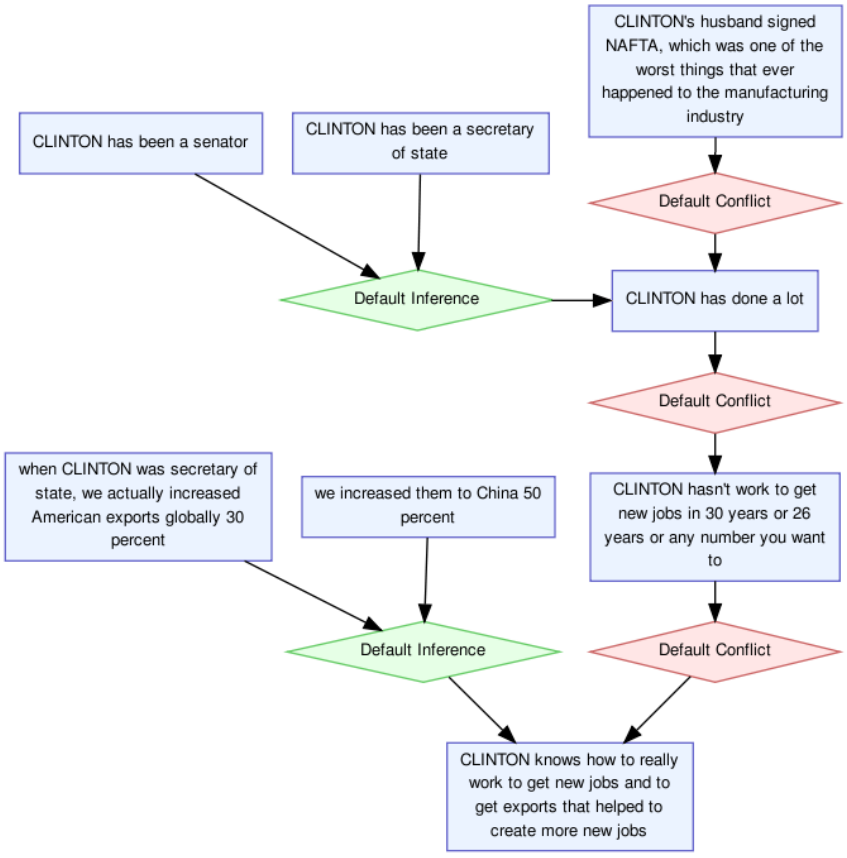


Figure 7.1: Fragment of Manually Analysed Argumentative Structure from the US2016G1tv Corpus.

In order to calculate centrality scores for each I-node, we adapt eigenvector centrality (used in the Google Pagerank algorithm (Brin and Page, 1998)).

Firstly, we create a directed graph, $G = (V, E)$, in which: the vertices (V) are propositions (I-nodes) extracted from the corpus; and, an edge exists between two vertices if there is an RA- or CA-node connecting them in the original analysis. We then construct the matrix $A = (a_{v,t})$, where $a_{v,t}$ is the weight of the edge between vertex v and vertex t if v and t are connected, and $a_{v,t} = 0$ otherwise. The relative centrality score of vertex v can then be defined as shown in Equation 7.1, where λ represents the greatest eigenvalue for which a non-zero eigenvector exists.

$$\text{Central}(v) = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} \text{Central}(t) \tag{7.1}$$

This results in a centrality score for each proposition, from which we can rank the propositions by how central they are to the debate. The top four ranked central propositions from US2016G1tv are listed below:

- CLINTON could encourage them by giving them tax incentives, for example
- there is/is not any way that the president can force profit sharing
- CLINTON also wants to see more companies do profit-sharing
- CLINTON is hinting at tax incentives

It is encouraging that these issues all concern the economy, which Pew Research identified as the single most important issue to voters (with 84% of voters ranking it as "very important") in the 2016 US presidential elections[1].

7.2.2 Divisiveness

Divisive issues are those that split opinion and which have points both supporting and attacking them (Konat et al., 2016). Looking again at Figure 7.1, we can see that the node "CLINTON knows how to really work to get new jobs..." is not only central, but also divisive, with both incoming support and conflict. At the opposite end of the scale, the node "CLINTON has been a

[1] `http://www.people-press.org/2016/07/07/4-top-voting-issues-in-2016-election/`

secretary of state", is not divisive; such factual statements are unlikely to be disputed by anyone on either side of the debate.

The Divisiveness of an issue measures how many others are in conflict with it and the amount of support which the two sides have. In order to calculate this, we now create two directed sub-graphs, one for support and one for conflict: $G_s = (V, E_s)$, in which an edge exists between two vertices if there is an RA-node connecting them in the original analysis; and, $G_c = (V, E_c)$, in which an edge exists between two vertices if there is a CA-node connecting them in the original analysis. The divisiveness of a vertex v can then be defined as shown in Equation 7.2, where $deg_s^-(v)$ is the in-degree of vertex v in G_s.

$$\text{Divisive}(v) = \sum_{\forall t \text{ s.t. } (t,v)\in E_c \vee (v,t)\in E_c} deg_s^-(v) * deg_s^-(t) \tag{7.2}$$

Again we list the top four ranked divisive issues from US2016G1tv below, and it is certainly easy to see how such statements on the character of the candidates, the validity of their claims and controversial issues such as gun control could easily divide those commenting on the debate:

- TRUMP settled that lawsuit with no admission of guilt
- I still support hand guns though
- people have looked at both of our plans, have concluded that CLINTON's would create 10 million jobs and TRUMP's would lose us 3.5 million jobs
- CLINTON didn't realize coming off as a snarky teenager isn't a good look either

7.3 Automating the Identification of Large Scale Argument Graph Properties

In this section we investigate techniques to automatically rank text fragments by their centrality and divisiveness with no prior knowledge of the argumentative structure contained within the text. In each case, we take the manually

segmented propositions from our corpus and apply techniques to rank these, we then compare the resulting rankings to the ranking determined from the manually analysed argument structures as described in Section 7.2.

7.3.1 Automatic Identification of Centrality

In order to calculate centrality automatically, we take advantage of the results from Chapter 6 which show that propositions (I-nodes) that are connected by relations of either support or attack in an AIF graph will generally have a higher (lexical or semantic) similarity than those which have no argumentative connection. We can again see an example of this in Figure 7.1, where the node "CLINTON knows how to really work to get new **jobs** and to get **exports** that..." is connected via support and attack relations to nodes whose propositional contents are all related to jobs or exports. The remaining nodes in this example fragment all discuss more distant concepts, such as Clinton's experience.

As such, centrality of propositions can be automatically calculated by determining the similarity scores between all proposition pairs and then computing eigenvector centrality on a graph with edge weights corresponding to these similarity scores. The resulting automatically calculated centrality scores should then mirror those that would be determined by calculating centrality on the argument graph.

We consider those methods for determining similarity shown to perform best in Chapter 6: ROUGE-1, ROUGE-2, ROUGE-L, Levenshtein Edit Distance, Word Vectors, and Document Vectors). And, for each of these measures, produce an automatically generated similarity graph, using the similarity score as the edge weights. For each of these graphs, the eigenvector centrality scores were then calculated, and the vertices sorted by centrality score to give an ordered list of propositions.

The ranking obtained using each centrality measure was then compared to the centrality ranking calculated for the manually annotated argument structure (as described in Section 7.2.1), by calculating the Kendall rank correlation coefficient (Kendall, 1938). The results for each method are shown in Table 7.1.

In each case the results show a correlation between the rankings ($p < 0.05$) suggesting that all of these methods are able to approximate the centrality of propositions in the argumentative structure. In Section 7.4 we explore these results further and show that these approximations are in all cases sufficient to improve the automatic extraction of the argumentative structure directly from the original text.

Similarity Method	**Kendall τ**
ROUGE-1	0.585
ROUGE-2	0.601
ROUGE-L	0.533
Levenshtein Edit Distance	0.524
Word vectors	0.618
Document vectors	0.620

Table 7.1: The Kendall rank correlation coefficient (τ) for the rankings determined using TextRank for each method of determining semantic similarity compared to the Centrality ranking obtained from the manually annotated argument structure.

7.3.2 Automatic Identification of Divisiveness

Whilst divisiveness is a related concept to centrality, it is more challenging to determine directly from the text, as we need to not only locate those nodes that are most discussed, but also to limit this to those which are involved in conflict relations.

Here we implement a method of determining conflict relations using SentiWordNet[2], a lexical resource for opinion mining[3]. SentiWordNet assigns a triple of polarity scores to each synset of WordNet, a positivity, negativity and

[2] `http://sentiwordnet.isti.cnr.it/`

[3] It should be noted that the state-of-the-art in sentiment analysis continues to move rapidly, and improved results than those presented here may be achieved with more recent techniques, such as XLnet (Yang et al., 2019)

$$Polarity(P_i, P_j) = \frac{|positivity(P_i) - positivity(P_j)| + |negativity(P_i) - negativity(P_j)|}{2} \tag{7.3}$$

objectivity score. The sum of these scores is always 1. For example, the triple (1, 0, 0) (positivity, negativity, objectivity) is assigned to the synset of the word "good".

Each proposition (I-node), is split into words and each word is stemmed and tagged, and stop words are removed. If a stemmed word belongs to one of the word classes "adjective", "verb" or "noun", its polarity scores are looked up in SentiWordNet. Where a word has multiple synsets, each of the polarity scores for that word are averaged across all of its synsets. The scores of all words within a sentence are then summed and divided by the number of words with scores to give a resulting triple of {positivity, negativity, objectivity} values for each proposition.

Having calculated the polarity triples for each proposition, we are then able to calculate the difference in polarity between two propositions, P_i and P_j as in Equation 7.3.

We compute these differences in polarity for each pair of propositions in the corpus and then, for each of the methods of determining similarity discussed in the previous subsection, multiply the similarity scores by the polarity difference to obtain a value representing the likelihood of conflict between the two. Finally for each proposition, we mirror the method of computing divisiveness from the argument graph. To do this, we look at each proposition, and take the sum of the centrality scores multiplied by the conflict value for each other proposition.

Following this approach for each method of determining similarity again gives us a ranking which we can then compare to the divisiveness ranking calculated for the manually annotated argument structure, as described in Section 7.2. For each approach, we again calculate the Kendall rank correlation coefficient. These results are shown in Table 7.2. We can see from these results that whilst there is still a positive correlation between the rankings, these are

substantially less significant than those obtained for the centrality rankings. In the next Section we investigate whether these values are sufficient to have a positive impact on the argument mining task.

Similarity Method	**Kendall τ**
ROUGE-1	0.113
ROUGE-2	0.237
ROUGE-L	0.147
Levenshtein Edit Distance	0.224
Word vectors	0.167
Document vectors	0.284

Table 7.2: The Kendall rank correlation coefficient (τ) for the Divisiveness rankings for each method of determining semantic similarity compared to the Divisiveness ranking obtained from the manually annotated argument structure.

7.4 Validation: Applying Automatically Identified Centrality and Divisiveness Scores to Argument Mining

Our final step is to validate both our concepts of centrality and divisiveness as calculated from annotated argument structures and our methods of calculating these same metrics directly from unannotated text. To do this, we use the "Topical Similarity" argument mining technique presented in Chapter 6 and in Lawrence et al. (2014).

Starting with the results for each similarity measure as given in Section 6.2, we here adapt Step 2 of this process by considering all of the previous propositions in the corpus as potential candidate structures and, having produced these candidate structures calculated the Centrality and Divisiveness rankings for each structure as described in Section 7.2. Finally we computed the Kendall

rank correlation coefficient comparing the centrality ranking of each candidate structure to the ranking computed only using similarity (as described in Section 7.3) and selected the structure which maximised the rank correlation.

Table 7.3 shows the precision, recall and F1-scores for automatically determining connections, based on attachment only, in the US2016G1tv corpus using each semantic similarity measure combined with maximising the rank correlations for centrality and divisiveness. We can see from these results that maximising divisiveness results in small increases in accuracy, and in all cases maximising centrality results in increased accuracy in determining connections, with increases of 0.03–0.10 in F1-score demonstrated for all the methods considered.

7.5 Conclusion

This chapter has presented two metrics, Centrality and Divisiveness, for describing the nature of propositions and their context within a large scale argumentative structure. We have shown how these metrics can be calculated from annotated argument structures and produced reliable estimations of these metrics that can be extracted directly from un-annotated text, with strong positive correlations between both rankings.

Finally, we have shown how these metrics can be used to improve the accuracy of existing argument mining techniques. By broadening the focus of argument mining from specific facets, such as classifying as premise or conclusion, to look at features of the argumentative structure as a whole, we have presented an approach which can improve argument mining results either as a feature of existing techniques or as a part of a more robust combined technique such as that presented in Chapter 9.

Similarity Method	**p**	**r**	**F1**
ROUGE-1	0.67	0.59	0.63
+ Max Centrality	**0.68**	**0.67**	**0.67**
+ Max Divisiveness	0.66	0.61	0.63
ROUGE-2	0.75	0.67	0.71
+ Max Centrality	**0.79**	**0.70**	**0.74**
+ Max Divisiveness	0.76	0.67	0.71
ROUGE-L	0.60	0.64	0.62
+ Max Centrality	**0.66**	**0.67**	**0.66**
+ Max Divisiveness	0.60	0.64	0.62
Levenshtein Edit Distance	0.62	0.68	0.65
+ Max Centrality	**0.67**	**0.71**	**0.69**
+ Max Divisiveness	0.63	0.70	0.66
Word vectors	0.73	0.63	0.68
+ Max Centrality	**0.77**	**0.67**	**0.72**
+ Max Divisiveness	0.75	0.65	0.70
Document vectors	0.71	0.60	0.65
+ Max Centrality	**0.75**	**0.66**	**0.70**
+ Max Divisiveness	0.70	0.63	0.66

Table 7.3: Precision, recall and F1-scores for automatically determining connections in the US2016G1tv corpus using each similarity measure combined with Centrality and Divisiveness.

Chapter 8

Argumentation Schemes

8.1 Introduction

Argumentation schemes (Walton, 1996) capture structures of (typically presumptive) inference from a set of premises to a conclusion and represent stereotypical patterns of human reasoning. As such, argumentation schemes represent a historical descendant of the topics of Aristotle (1958) and, much like Aristotle's topics, play a valuable role in both the construction and evaluation of arguments.

Several attempts have been made to identify and classify the most commonly used schematic structures (Hastings, 1963; Perelman and Olbrechts-Tyteca, 1969; Kienpointner, 1992; Pollock, 1995; Walton, 1996; Grennan, 1997; Katzav and Reed, 2004; Walton et al., 2008). Although these sets of schemes overlap in many places, the number of schemes identified and their granularity can be quite different. As such, most argument analyses tend to contain examples from only one scheme set, with the Walton set being the most commonly used. Several examples of Walton's argumentation schemes can be seen in Table 8.2.

Understanding the argumentative structure being expressed in a piece of natural language text can help us gain a deeper understanding of what is being said compared to many existing techniques for extracting meaning. If we consider the product review shown in Example (1), then sentiment analysis techniques allow us to understand at a high level what views are being pre-

sented, for example, that this review is positive, but are unable to provide details on exactly why the reviewer likes the product.

(1) The PowerShot SX510 is a fantastic camera. It is made by Canon and all Canon cameras have great image stabilisation.

Looking at the argumentative structure contained within this review, we can see that the propositions "It is made by Canon" and "all Canon cameras have great image stabilisation" are working together as a linked argument (see Section 2.2.3) to support the conclusion "The PowerShot SX510 is a fantastic camera". Furthermore, we can see that the link between the premises and conclusion is a form of Verbal Classification[1]. A graphical representation of the argument structure can be seen in Figure 8.1.

As shown in the examples in Table 8.2, Walton's classification assigns a particular label to each component part of a scheme instance. For the Verbal Classification in Example (1), the scheme components are shown below:

Premise (ContainsProperty): It is made by Canon

Premise (ClassificationProperty): all Canon cameras have great image stabilisation

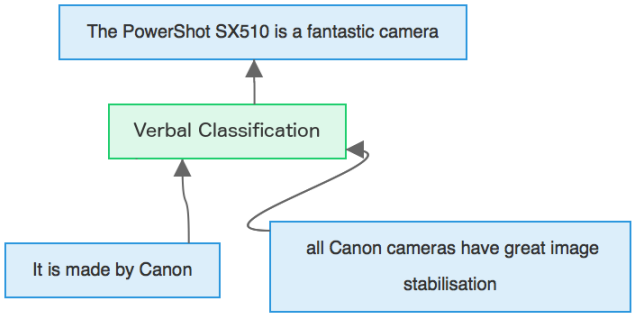
Conclusion: The PowerShot SX510 is a fantastic camera

The features of these common patterns of argument provide us with a way in which to both identify that an argument is being made and determine its structure. By using the specific nature of each component proposition in a scheme, we can identify where a particular scheme is being used and classify the propositions accordingly, thereby gaining a deeper understanding of the argumentative structure which a piece of text contains.

The concept of automatically identifying argumentation schemes was first discussed by Walton (2011) and Feng and Hirst (2011). Walton proposes a six-stage approach to identifying arguments and their schemes. The approach

[1]In fact, the example here does not exactly conform to the Verbal Classification scheme. In a more thorough analysis, an enthymeme would be added showing that the premises actually support the fact that the camera has great image stabilisation and that this in turn is a feature of a fantastic camera.

Figure 8.1: Argument analysis of a product review, showing an example of the Verbal Classification scheme



suggests first identifying the arguments within the text and then fitting these to a list of specific known schemes. A similar methodology was implemented by Feng & Hirst, who produced classifiers to assign pre-determined argument structures as one in a list of the most common argumentation schemes. Another possible approach is suggested in Cabrio et al. (2013), where the connection between argumentation schemes and discourse relations is highlighted, however, this requires these discourse relations to be accurately identified before scheme instances can be determined.

The main challenge faced by these approaches is the need for some prior analysis of the text to have taken place. By instead looking at the features of each component part of a scheme, we are able to overcome this requirement and identify parts of schemes in completely unanalysed text. Once these scheme components have been identified, we are able to group them together into specific scheme instances and thus obtain a complete understanding of the arguments being made.

8.2 Walton's Classification of Argumentation Schemes

As the starting point for our annotation of argument schemes based on Walton's typology, we use the collection in the book "Argumentation Schemes" by Walton et al. (2008). Depending on what is counted as a type of argument

scheme (i.e. whether sub-types are counted or not), the book contains upwards of 60 schemes. The schemes are presented with their distinctive pattern of premises and conclusion, and with an associated list of critical questions, mostly drawn from Walton's previous work.

8.2.1 Annotation Guidelines

Two expert annotators trained in argumentation analysis and with prior knowledge of Walton's typology of argument schemes each classified 55% of the RA-nodes in the US2016G1tv corpus in accordance with Walton's typology. Specifically, the top level schemes from (Walton et al., 2008) were considered, resulting in a choice from 60 possible labels to be applied to each of the more than 500 previously analysed inference relations in the corpus.

To facilitate the process, the annotators were provided with a classification decision tree: a heuristic for the annotators, to intuitively support their coding task (Lawrence et al., 2019b). The fragment of the heuristic in Figure 8.2 shows the indication of the grounds for making a decision between various action-oriented argument schemes. The decision tree ties into the actual guidelines consisting of Chapter 9 of (Walton et al., 2008, pp. 308–346): *A User's Compendium of Schemes*. Since the annotation relies on the existing annotated argumentation structure, in some cases, the schemes are applied in a simplified, condensed or partial manner, to fit the original annotation. In addition, one auxiliary catch-all class is introduced for arguments not fitting any of the 60 main schemes: *Default inference*.

8.2.2 Results of the Annotation

A sample of 10.2% of the corpus was annotated by both annotators, resulting in a Cohen's κ (Cohen, 1960) of 0.723; well within substantial agreement (Landis and Koch, 1977). A confusion matrix showing the results of the double annotation can be seen in Figure 8.3[2].

[2]Schemes which were not used by either annotator in the double annotation are omitted from the matrix

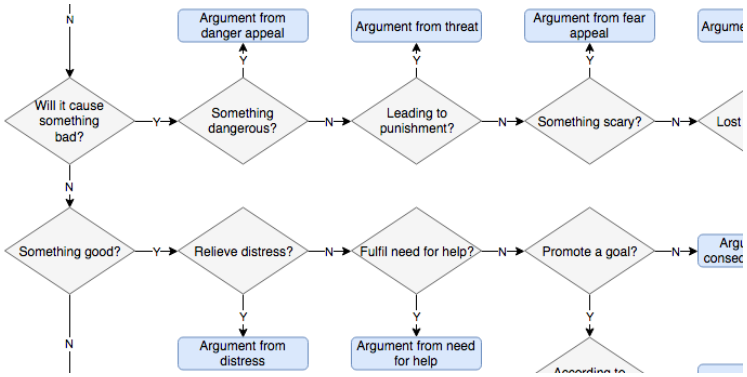


Figure 8.2: Distinguishing between action-oriented argument schemes with the decision tree heuristic.

Some classes of argument scheme turned out to be particularly difficult to distinguish: e.g., Example (2) was classified by one annotator as *Practical reasoning*, related to promoting goals, and by the other as *Argument from values*, related to promoting values.

(2) Hilary Clinton: *What I have proposed would be paid for by raising taxes on the wealthy [...] I think it's time that the wealthy and corporations paid their fair share to support this country.*

The results of the annotation in accordance with Walton's classification of argument schemes are collected in the US2016G1tvWALTON corpus (available online at `http://corpora.aifdb.org/US2016G1tvWALTON`). Figure 8.4 shows an example of the *Practical reasoning from analogy* scheme as applied in the corpus. Of the 505 RA-nodes in the original US2016G1tv corpus, a total of 491 are annotated with one of the 60 argument scheme types in Walton's classification, leaving only 14 as *Default inference*. The most common scheme, by some margin, is *Argument from example*. The *Argument from expert opinion* scheme, a scholarly favourite, is remarkably rare with only three occurrences. Table 8.1 shows the number of occurrences of each scheme within the corpus.

Annotator B \ Annotator A	Argument from alternatives	Argument from analogy	Argument from bias	Argument from cause to effect	Argument from consequences	Argument from example	Argument from expert opinion	Argument from fear appeal	Argument from oppositions	Argument from popular opinion	Argument from popular practice	Argument from sign	Argument from values	Argument from verbal classification	Argument from witness testimony	Contextual ad hominem	Default Inference	Generic ad hominem	Practical reasoning	Practical reasoning from analogy	Pragmatic argument from alternatives	Pragmatic inconsistency
Argument from alternatives	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Argument from analogy	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Argument from bias	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Argument from cause to effect	0	0	0	**4**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Argument from consequences	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0
Argument from example	0	0	0	0	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Argument from expert opinion	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Argument from fear appeal	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Argument from oppositions	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
Argument from popular opinion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Argument from popular practice	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0
Argument from sign	0	0	0	**2**	0	**1**	0	0	0	0	0	**7**	0	**2**	0	0	0	0	0	0	0	0
Argument from values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0
Argument from verbal classification	0	0	0	0	0	**1**	0	0	0	0	0	0	0	**4**	0	0	0	0	0	0	0	0
Argument from witness testimony	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0
Contextual ad hominem	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Default Inference	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0
Generic ad hominem	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	**2**	0	0	0	0
Practical reasoning	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**6**	0	0	0
Practical reasoning from analogy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0
Pragmatic argument from alternatives	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pragmatic inconsistency	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**

Figure 8.3: Confusion matrix for annotation of schemes in US2016G1tvWALTON

8.3 Automatic Identification of Argumentation Schemes

Being able to determine the argumentation scheme structure contained within a piece of text gives us a much deeper understanding of both what views are being expressed and why those views are held, as well as providing a route to the automatic reconstruction of certain types of enthymeme (Hitchcock, 1985). However, existing approaches to automatically identifying scheme instances have relied on the basic argumentative structure being previously identified.

By training a range of classifiers to identify the individual components of a scheme, we are able to identify not just the presence of a particular scheme, but also the roles which each of the premises play within a particular scheme

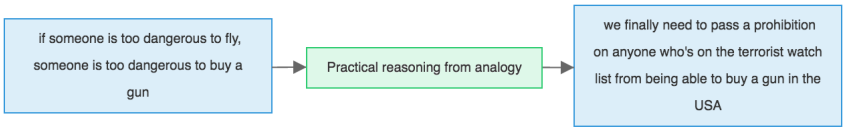


Figure 8.4: OVA visualisation of *Practical reasoning from analogy* in US2016G1tvWALTON.

Table 8.1: Counts of argument schemes in the US2016G1tvWALTON corpus.

Argument scheme	Count
Argument from example	81
Argument from cause to effect	48
Practical reasoning	45
Argument from consequences	40
Argument from sign	38
Argument from verbal classification	32
Generic ad hominem	28
Circumstantial ad hominem	24
Pragmatic argument from alternatives	23
Argument from values	15
Default inference	14
Argument from position to know	13
Argument from fear appeal	11
Argument from alternatives	9
Argument from bias	9
Argument from analogy	8
Argument from popular opinion	8
Argument from danger appeal	7
Argument from popular practice	7
Argument from composition	6

Argument scheme	Count
Ethotic argument	5
Practical reasoning from analogy	4
Argument from commitment	3
Argument from expert opinion	3
Argument from waste	3
Argument from gradualism	2
Argument from need for help	2
Argument from oppositions	2
Argument from perception	2
Argument from correlation to cause	1
Argument from definition to verbal classification	1
Argument from division	1
Argument from ignorance	1
Argument from rules	1
Argument from vagueness of verbal classification	1
Argument from witness testimony	1
Argumentation from interaction of act and person	1
Pragmatic inconsistency	1
Two-person practical reasoning	1

instance. Furthermore, we are able to perform this based only on a list of the propositions contained within the text, requiring no previous analysis to have been performed.

Limiting the data to those schemes with at least thirty instances that are fully defined leaves us with six schemes to consider (comparable in number to the top five most commonly occurring schemes used by Feng and Hirst (2011)): Argument from example, Argument from cause to effect, Practical reasoning, Argument from consequences, Argument from sign, and Argument from verbal classification. The structure of these schemes, and their associated component types, is shown in Table 8.2.

In Section 8.3.1 we look at using one-against-others classification to identify propositions of each scheme component type from the set of propositions in US2016G1tv. Being able to successfully perform this task for even one of the proposition types allows us to discover areas of the text where the corre-

Argument from example (EX) *Premise (Pr):* In this particular case, the individual a has a property F and also property G. *Conclusion (Cn):* Therefore, generally, if x has property F, then it also has property G.
Argument from cause to effect (CE) *Major Premise (Mj):* Generally, if A occurs, then B will (might) occur *Minor Premise (Mn):* In this case, A occurs (might occur) *Conclusion (Cn):* Therefore, in this case, B will (might) occur
Practical reasoning (PR) *Major Premise (Mj):* I have a goal G *Minor Premise (Mn):* Carrying out this action A is a means to realise G. *Conclusion (Cn):* Therefore, I ought (practically speaking) to carry out this action A.
Argument from consequences (CS) *Premise (Pr):* If A is brought about, then good/bad consequences will plausibly occur. *Conclusion (Cn):* Therefore, A should/should not be brought about.
Argument from sign (SN) *Specific Premise (Sp):* A (a finding) is true in this situation. *General Premise (Ge):* B is generally indicated as true when its sign, A, is true. *Conclusion (Cn):* B is true in this situation.
Argument from verbal classification (VC) *Individual Premise (In):* a has property F. *Classification Premise (Cl):* For all x, if x has property F, then x can be classified as having property G. *Conclusion (Cn):* a has property G.

Table 8.2: Examples of Walton argumentation schemes

sponding scheme is likely to be being used.

In order to accomplish these tasks, a range of classifiers for each scheme component type (all premise types, and the conclusion for each of the six schemes) was implemented using the *scikit-learn*[3] Python module for machine learning, with the features described in Table 8.3. Part Of Speech (POS) tagging was performed using the Python NLTK[4] POS-tagger and the frequencies of each tag added as individual features. The similarity feature was added to extend the information given by unigrams to include an indication of whether a proposition contains words similar to a pre-defined set of keywords. The keywords used for each type are shown in Table 8.4. Similarity scores were calculated using WordNet[5] to determine the maximum similarity between the synsets of the keywords and each word in the proposition. The maximum score for the words in the proposition was then added as a feature value, indicating the semantic relatedness of the proposition to the keyword.

8.3.1 One-against-others scheme component classification

For each of the schemes in Table 8.2, the conclusions and each type of premise were classified using three different types of classifier (Multinomial Naïve Bayes, Support Vector Machines (SVMs) and Decision Trees) against a random, equally sized, selection of other argument propositions from the US2016G1tv corpus.

Table 8.5 shows the precision, recall and F-score obtained using 10-fold cross validation for each proposition type with each classifier. For each proposition type, the F-Score of the best performing classifier is highlighted in bold.

As can be seen from the table, the Multinomial Naïve Bayes classifiers perform best in most cases, and even for those proposition types where one of the other methods perform better, the results are comparable. In particular, the results for SVMs are lower than those for the other types of classifier. This

[3] `http://scikit-learn.org/stable/`

[4] `http://www.nltk.org/`

[5] `http://wordnet.princeton.edu/`

Feature	**Description**
Unigrams	Each word in the proposition
Bigrams	Each pair of successive words
Length	The number of words in the proposition
AvgWLength	The average length of words in the proposition
POS	The parts of speech contained in the proposition
Punctuation	The presence of certain punctuation characters, for example “ ” indicating a quote
Similarity	The maximum semantic similarity of a word in the proposition to predefined words corresponding to each proposition type calculated based on the distance between their WordNet synsets

Table 8.3: Features used for classification

can be explained by the fact that our feature set is considerably larger than the sample, a situation in which SVMs generally perform less well.

Notably, the results for Argument from Example (Premise) and Argument from Verbal Classification (Conclusion) are quite weak in comparison to the other proposition types. In the case of Argument from Example, although the premise component description (“In this particular case, the individual a has a property F and also property G”) seems quite specific, actual instances of this scheme component in the annotated corpus are less clear. E.g. “one needs more police” is given as an example supporting “one needs a better community”, and “Ford is leaving” given as an example supporting “Thousands of jobs leaving Michigan , leaving Ohio”. In both of these cases, the premise takes the form of

Proposition	**Keywords**
EX Pr	example
EX Cn	generally
CE Mj	generally, occurs
CE Mn	occurs
CE Cn	occurs
PR Mj	goal
PR Mn	action, doing
PR Cn	should, perform
CS Pr	result, outcome, good, bad
CS Cn	should, ought
SN Sp	situation, this, here
SN Ge	generally, true, case
SN Cn	situation, this, here
VC In	property, is
VC Cl	property, also, similarly
VC Cn	property, is

Table 8.4: Keywords used for each scheme component type

very simple statement (policy based in the first case and factual in the second case), making them hard to distinguish – it is almost impossible to highlight a difference between "Ford is leaving" which *is* used as an example, and other segments such as "Rahami is still alive" which is *not* used as an example.

With Argument from Verbal Classification, the issue is perhaps clear even from the conclusion component description ("a has property G"). This means that simple fact based statements such as "it's a big problem" and "CLINTON is wrong" fall into this category, and are almost impossible to separate from similar statements which do not.

The results for the remaining proposition types are more promising and, even for those schemes where the classification of one proposition type is less successful, the results for the other types are better. If we consider being able

Type	**Naïve Bayes**			**SVM**			**Decision Tree**		
	p	r	f1	p	r	f1	p	r	f1
EX Premise	0.53	0.53	**0.53**	0.48	0.46	0.47	0.49	0.52	0.50
EX Conclusion	0.69	0.65	**0.67**	0.60	0.67	0.63	0.63	0.64	0.63
CE Major Premise	0.80	0.70	**0.75**	0.60	0.62	0.61	0.76	0.70	0.73
CE Minor Premise	0.61	0.79	**0.69**	0.58	0.51	0.54	0.58	0.79	0.67
CE Conclusion	0.68	0.65	0.66	0.56	0.60	0.58	0.63	0.73	**0.68**
PR Major Premise	0.77	0.73	0.75	0.68	0.80	0.74	0.64	0.93	**0.76**
PR Minor Premise	0.68	0.80	**0.74**	0.64	0.71	0.67	0.75	0.64	0.69
PR Conclusion	0.80	0.96	**0.87**	0.76	0.85	0.80	0.89	0.83	0.86
CS Premise	0.65	0.75	**0.70**	0.71	0.64	0.67	0.65	0.68	0.66
CS Conclusion	0.78	0.73	0.75	0.71	0.74	0.72	0.76	0.82	**0.79**
SN Specific Premise	0.67	0.79	0.73	0.61	0.61	0.61	0.76	0.75	**0.75**
SN General Premise	0.84	0.76	**0.80**	0.69	0.65	0.67	0.74	0.75	0.74
SN Conclusion	0.58	0.6	**0.59**	0.46	0.5	0.48	0.48	0.59	0.53
VC Indiv. Premise	0.71	0.74	**0.72**	0.45	0.88	0.60	0.57	0.92	0.70
VC Class. Premise	0.69	0.80	**0.74**	0.72	0.66	0.69	0.66	0.61	0.63
VC Conclusion	0.58	0.49	0.53	0.40	0.47	0.43	0.50	0.58	**0.54**

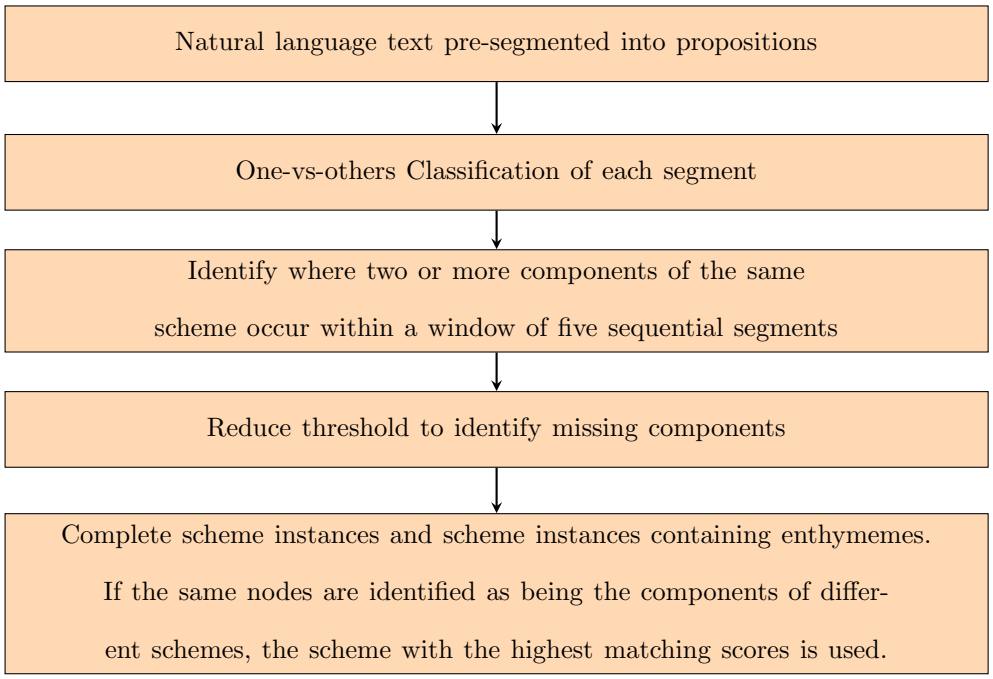
Table 8.5: Results of one vs others proposition classification using 10-fold cross validation (The highest f-score for each scheme component is highlighted in bold)

to correctly identify at least one proposition type, then our results give F-scores between 0.67 and 0.87 for locating an occurrence of the different scheme types. The results also show that in many cases it would be possible to not only determine that a scheme is being used, but to accurately classify all of its component propositions.

8.3.2 Identification of Scheme Instances

The one-against-others results suggest that it is feasible to classify propositions by type. Performing this classification on a piece of text would enable us to identify places where a particular scheme is being used. We now move on

Figure 8.5: Process used for identifying scheme instances from segmented text



to look at how well these classifiers are able to identify not just individual occurrences of a proposition type but complete scheme instances. The ability to successfully perform this task would enable us to take a sample of natural language and understand a large amount of the argument structure it contains.

The aim of this experiment is not to identify the complete argumentative structure represented by the text, but to illustrate that it is possible to use the classifiers that we have produced to extract complete scheme instances. In order to accomplish this, we first perform one-vs-others classification of each segment using the Multinomial Naïve Bayes classifiers discussed in Section 8.3.1. We then look at each group of five sequential segments, and identify places where two or more components of the same scheme type occur together. In cases where there is still a missing component, we reduce the threshold for the classifier corresponding to the missing piece. If reducing the threshold still does not offer a candidate for the missing scheme component, we assume that this is unstated enthymematic content in the argument. By performing these steps, we are able to take segmented text and identify either complete scheme instances, or partial scheme instances which have some enthymematic component. The process followed is illustrated in Figure 8.5.

This classification process identified 46 possible occurrences of Argument from example, 17 of Argument from cause to effect, 22 of Practical reasoning, 35 of Argument from consequences, 18 of Argument from sign, and 9 of Argument from verbal classification.

Figure 8.6: Automatically identified Argument from Consequences instance

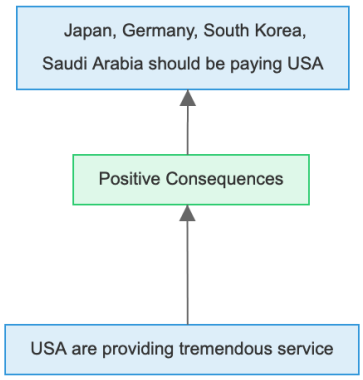
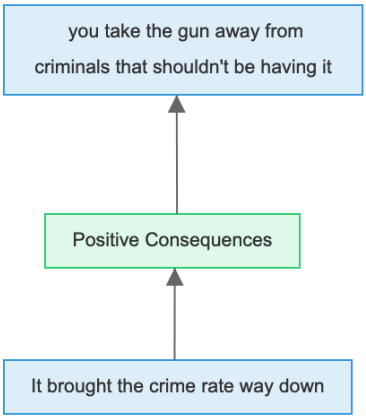


Figure 8.7: Partially correct automatically identified Argument from Consequences instance

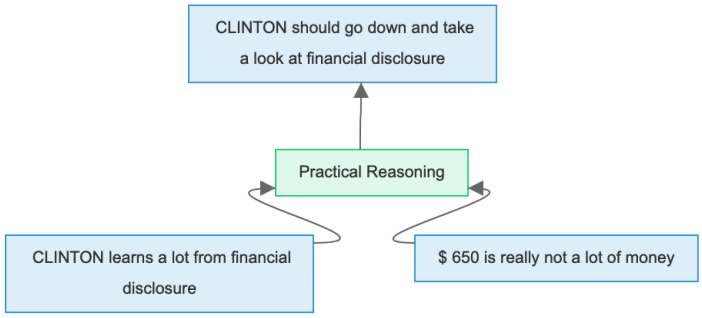


An example of a correctly identified instance of Argument from Consequences can be seen in Figure 8.6. However, the instance of Argument from Consequences shown in Figure 8.7 is only partially correct. In this case, the premise, "It brought the crime rate way down" matches with the gold standard annotation, and fits the criteria for good consequences, but, the identified conclusion "you take the gun away from criminals that shouldn't be having it" is unconnected in the annotation, and should actually be "whether or not

in a place like Chicago you do stop and frisk, which worked very well, Mayor Giuliani is here, worked very well in New York".

A final example, this time showing a partially correct identified instance of Practical reasoning, is shown in Figure 8.8. In this case, the conclusion ("CLINTON should go down and take a look at financial disclosure") and the minor premise ("CLINTON learns a lot from financial disclosure") are correct, with the former matching an action and the latter this action realising a goal. However, in this case, the major premise is actually left implicit according to the annotation, and has been incorrectly identified as "$650 is really not a lot of money".

Figure 8.8: Partially correct automatically identified Practical reasoning instance



Although these examples are not perfect identifications of scheme instances, it is clear that even with the limitations involved, we have come close to being able to identify at least where a scheme is occurring, and to correctly assign at least some of the propositions.

8.4 Conclusion

Whilst argumentation schemes have been detailed extensively in philosophy and psychology, perhaps due to the relative complexity of these structures, they have received little attention in argument mining. In Feng and Hirst (2011), instances of particular schemes are classified from text which has previously been annotated for its argumentative structure, a process which could

be considered as the second step in the six-stage approach to identifying arguments and their schemes suggested by (Walton, 2011).

Here, we have shown that by considering the features of the individual types of premise and conclusion that comprise a scheme, it is possible to classify these scheme components with reasonable accuracy. Despite the differing goals, our results are comparable results to those of Feng & Hirst, where the occurrence of a particular argumentation scheme was identified with accuracies of between 62.9% and 90.8%. Our results show that, on the same dataset, it is possible to identify individual scheme components with similar performance (F-scores between 0.67 and 0.87) can be achieved in identifying argumentation schemes in unanalysed text.

Furthermore, by searching for groupings of these proposition types, we have shown it is possible to determine not just that a particular scheme is being used, but to correctly assign propositions to their schematic roles. In future work accuracy of these techniques could be further improved by an in-depth review of the features used for classification using an ablation study, and by considering domain specific schemes, such as the Consumer Argumentation Scheme (CAS) (Wyner et al., 2012) aimed specifically at product reviews.

Chapter 9

A Combined Explainable Approach

9.1 Introduction

In the previous chapters (Chapter 4 to Chapter 8), five techniques for automatically determining information about the argumentative structure of a piece of text have been introduced. These techniques each draw inspiration from a different facet of the complex way in which humans understand the structure of an argument, and in doing so, provide explainable reasoning for the argumentative connections made. The individual techniques presented have been shown to produce capable results on their own, but, much as the theoretical works on which they are based each illuminate different and complementary aspects of human understanding, it is in their combination that a more full and accurate picture emerges. This chapter explores the way in which the previously presented techniques can be brought together, testing the resulting combined approach on the same US2016G1tv corpus as has been previously used for each individually.

In order to maintain explainability, a rule-based approach to combining is employed here. It would also be possible to use a machine learning approach for combination, with the different techniques being used as features, however this would then remove the ability to say which of the techniques was responsible for the classification. As the machine learning results would be based

on more complex interactions of the different techniques, it would be difficult to say exactly which were responsible for the classification. Machine learning combinations are however implemented for the purposes of evaluation (see Section 9.6), and, in order to provide a full comparison to broader work in the field, the combined approach is further tested on two additional widely used corpora (the Argument Annotated Essays corpus (Stab and Gurevych, 2017), and the Argumentative Microtext corpus (Peldszus and Stede, 2016)).

9.2 Combining the XAM Techniques

Each of the previously introduced techniques have displayed strengths and weaknesses when used in isolation, illuminating individual parts of the argumentative structure, yet falling short of providing a full and detailed picture. Discourse Indicators pick up simple linguistic cues, and have been shown to be very reliable in determining argumentative relations when they occur, though are not present in ∼80% of cases (see Chapter 4 and (Lawrence and Reed, 2015)). Premise-Conclusion Topic Models (Chapter 5), represent a listener's background knowledge concerning common themes of inference and can give a good indication of support relations, however on their own they are not specific enough to identify when these relations occur and when similar topics are being discussed without any inferential intent. Similarity measures (Chapter 6) have been shown to correlate significantly with argumentative relations, however they do not give an indication of the direction or type of relationship. Graph Properties (Chapter 7), as with similarity measures, give valuable clues about the argumentative structure (for example, a highly central node is likely to be a conclusion or main claim), however are not enough to provide structural detail on their own. Finally, Argument Scheme component detection (Chapter 8) allows common patterns of reasoning to be captured and labelled, but does not provide these labels in cases where such patterns are only loosely followed, or where there is insufficient training data for a particular scheme.

In order to maintain explainability, a rule-based approach to combination is adopted. Machine learning approaches could also be used to perform the

combination with the results from each technique being employed as features, however these would lose the ability to say clearly which of the individual techniques is being used to assign the labelling (for example to say that the system believes there is an inference relation between x and y because they form an instance of a particular argument scheme, or between y and z because of the presence of a discourse indicator). Section 9.6 contains a brief exploration of such machine learning approaches and shows that not only is explainability lost, but the overall results are also weaker.

The rule-based approach applies the techniques in order, building the full argumentative structure based on the parts of it that can be best identified by each technique. This rule-based combination method aims to leverage the strengths of each individual technique, whilst minimising their weaknesses. The rules followed are listed in order below:

1. **Discourse Indicators:** For every consecutive pair of ADUs, a and b, if the discourse indicator "because" exists between a and b, then, an RA with edges from b and to a is added. Discourse indicators are applied first as, despite their relatively rare use in real-world text, they have the highest precision of all the techniques.

2. **Argument Schemes:** If all components of a scheme are found (with a probability >80% for each) within a sliding window of five ADUs, then an RA is added linking the premise components to the conclusion components, and the RA is labelled with the identified scheme. This technique is used early on in the process as, when all scheme components are found together, we can say with confidence that an instance of that scheme is being used.

3. **Similarity:** For every consecutive pair of ADUs (including pairs that are split by a turn boundary), a and b, if the adjacency similarity threshold of any similarity measure is exceeded, then, an untyped and undirected edge connecting a and b is added.

4. **Similarity:** For every pair of ADUs, a and b, if the long distance similarity threshold of any similarity measure is exceeded, then, an untyped

and undirected edge connecting a and b is added. The two similarity steps aim to fill in the remaining edges in the argument structure before typing and direction is added by consideration of Premise-Conclusion Topic Models and Graph Properties.

5. **Premise-Conclusion Topic Models:** For each untyped edge added so far, if the TopicDist score (see Chapter 5) in either direction is greater than the mean topical inference matrix value, then an RA-node is added to the edge and the relevant direction assigned.

6. **Graph Properties:** For each remaining untyped and undirected edge, the direction is determined as being from the least central ADU to the most central, and an RA is added between them.

The rules start by searching for discourse indicators between ADUs. This method has been shown to be extremely reliable for certain indicators when they are present in the text, with the specific indicator used here, "because", shown in Chapter 4 to achieve a precision of 0.873. These connections are added first and viewed as being correct from then on.

The next step looks at argument schemes, and specifically those cases where all scheme components are identified within close proximity. Looking only at those cases where all components are found gives high confidence that this is indeed an instance of that particular scheme, and as such, the RA-node and scheme label are added connecting the components.

Similarity measures are used in steps 3 and 4 to fill in remaining connections, though this technique does not provide directionality or type for these edges. To fill in these details, premise-conclusion topic models and graph properties are used in steps 5 and 6. In step 5, the topics for each ADU at the end of an undirected edge are calculated and if the topical inference matrix shows that there is a high likelihood of inference in either direction, an RA-node is added in the appropriate direction. For any remaining undirected edges, the centrality scores for the ADUs at each end are calculated and there is assumed to be an inference relation going from the least central to the most central.

Details of which rule has been triggered to create each of the resulting relations are stored using the methodology described in Section 9.4.

9.3 Results

The result of this step-by-step rule application for an excerpt from US2016G1tv (Figure 9.1) is shown in Figures 9.2, 9.3, and 9.4. The bottom right graph shows the gold standard annotation for this text.

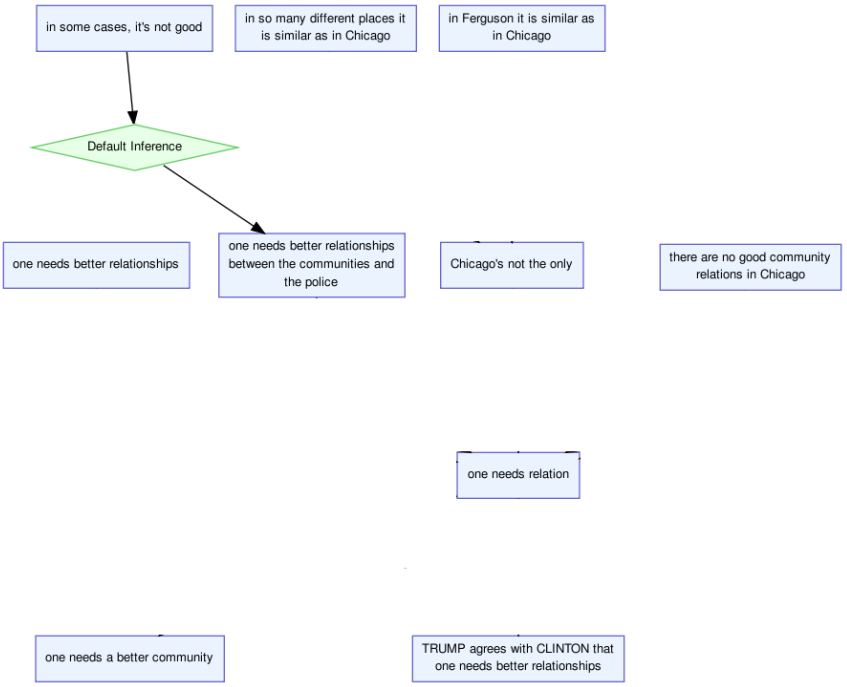
> **TRUMP:** You need a better community, you know, relation. You don't have good community relations in Chicago. But when you look—and Chicago's not the only—you go to Ferguson, you go to so many different places. You need better relationships. I agree with Secretary Clinton on this. You need better relationships between the communities and the police, because in some cases, it's not good.

Figure 9.1: An excerpt from the US2016G1tv corpus (map 10850 (http://www.aifdb.org/argview/10850))

In step 1, the discourse indicator "because" is found between "You need better relationships between the communities and the police" and "in some cases, it's not good", and an RA-node is added from the latter to the former. Step 2 identifies the inferential connection between "Chicago's not the only" and "in so many different places it is similar as in Chicago", however, the scheme components identified are those of argument from example, with "Chicago's not the only" being identified as the example (premise), and, as such, the inference is added in the wrong direction.

Steps 3&4 add a number of undirected edges based on the similarity scores of both adjacent and long range pairings. At this stage, the majority of the connections are added and although there are some changes to the structure, many of these are still valid. For example, although "TRUMP agrees with CLINTON that one needs better relationships" is now connected to "one needs better relationships" a rephrased version of the ADU it is connected to in the gold standard ("one needs relation").

In step 5, inference from "one needs better relationships between the communities and the police" to "TRUMP agrees with CLINTON that one needs better relationships", is added based on a tendency for the topics in the former to support those in the latter. Finally, step 6 adds types and directions to the



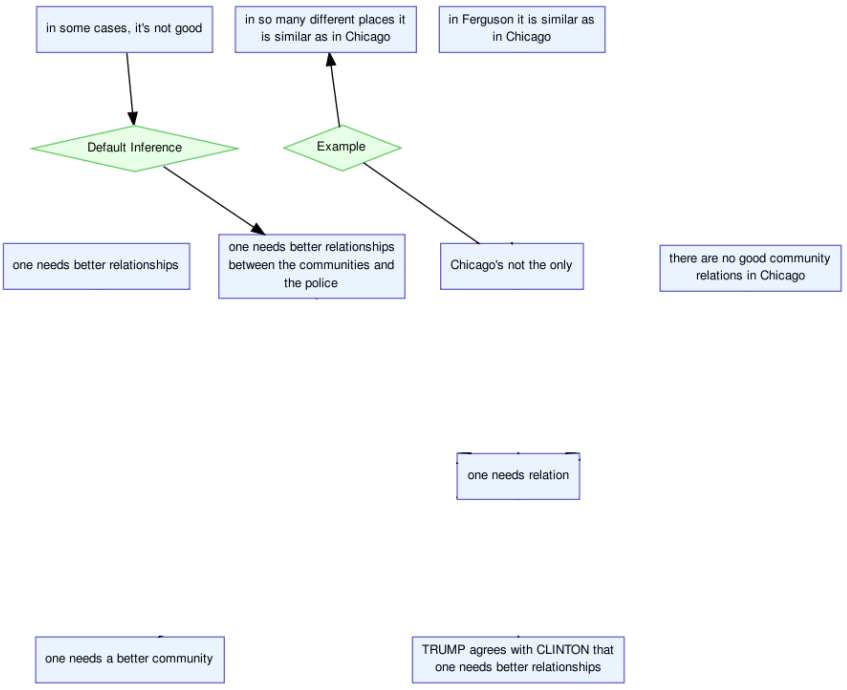


Figure 9.2: The result after steps 1 and 2 of rule based combination working on an excerpt from US2016G1tv

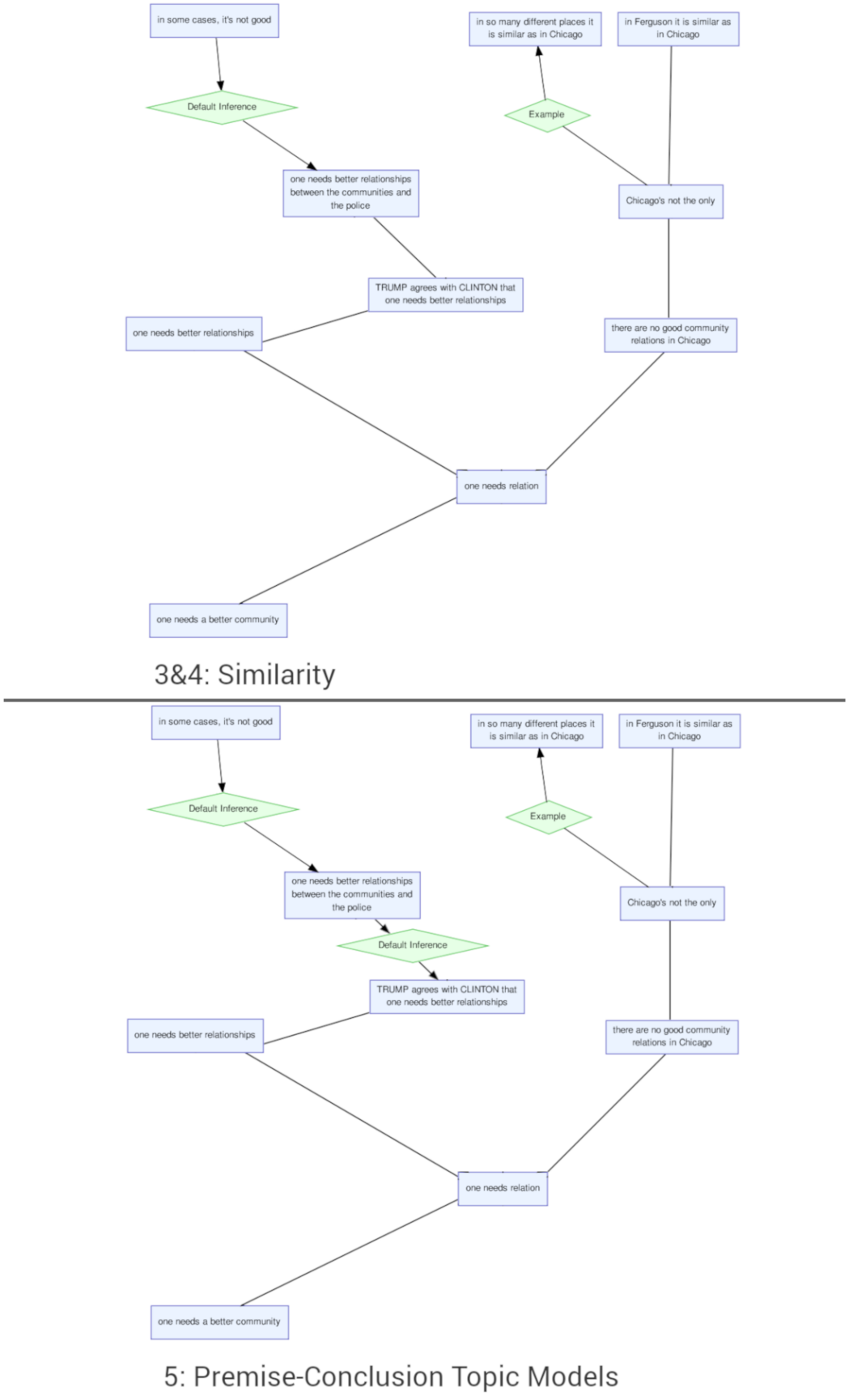
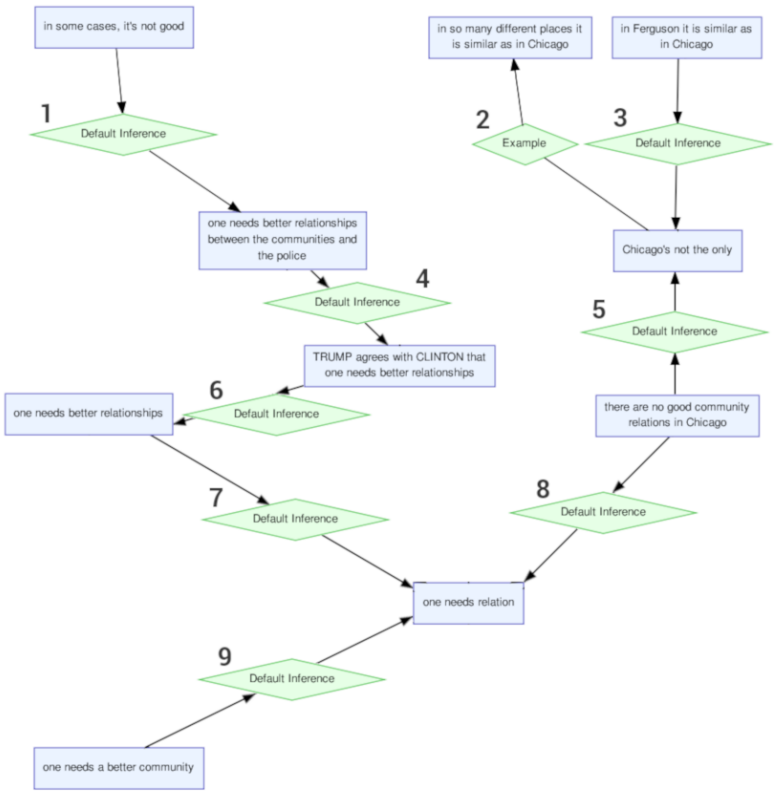


Figure 9.3: The result after steps 3, 4 and 5 of rule based combination working on an excerpt from US2016G1tv



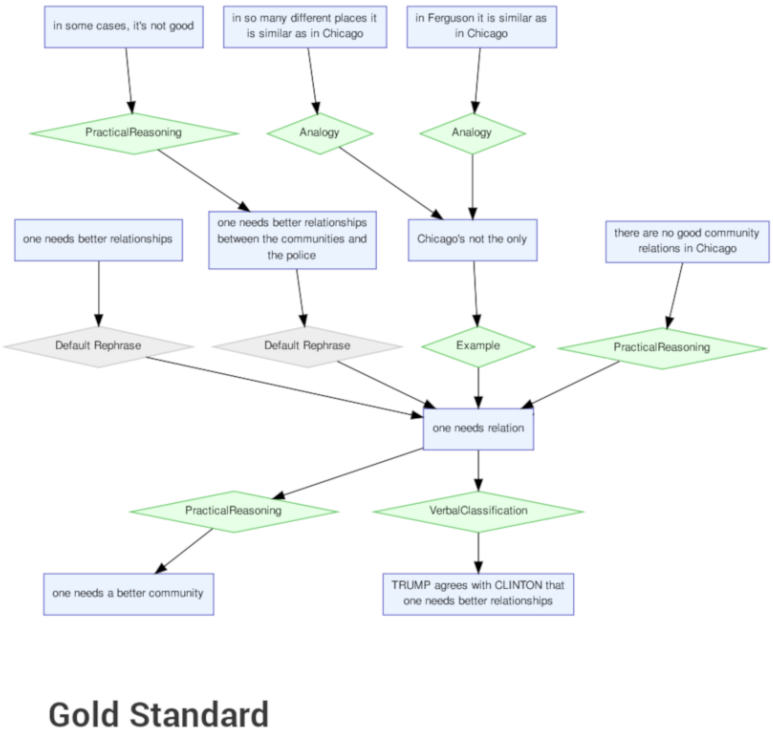


Figure 9.4: The result after step 6 of rule based combination working on an excerpt from US2016G1tv compared to the gold standard annotation

remaining edges. For example, "one needs relation" is determined as being more central than those nodes surrounding it, and edges are added coming into this.

The interpretation of classifications for each of the nine relations added in this example (as numbered in the top half of Figure 9.4) are given below:

1. Connection: **correct**; Direction: **correct**

2. The wrong scheme is assigned, and the direction is incorrect, however these two nodes are connected.
   Connection: **correct**; Direction: **incorrect**

3. Connection: **correct**; Direction: **correct**

4. The ADU "one needs better relationships between the communities and the police" is a rephrasing of "one needs relation" which supports "TRUMP agrees with CLINTON that one needs better relationships", hence this is viewed as correct (see Chapter 3 for a description of how rephrase relations are handled).
   Connection: **correct**; Direction: **correct**

5. Connection: **incorrect**; Direction: **incorrect**

6. Again, "one needs better relationships" is a rephrasing of "one needs relation" though the direction is, in this case, incorrect.
   Connection: **correct**; Direction: **incorrect**

7. The rephrase relation here has been mislabelled as inference,
   Connection: **incorrect**; Direction: **incorrect**

8. Connection: **correct**; Direction: **correct**

9. Connection: **correct**; Direction: **incorrect**

As can be seen from the above, there are several cases where an inference relation is correctly identified between two ADUs, but the direction of the inference is reversed. For this reason, the results shown in Table 9.1, show precision, recall and F-score for both directed and undirected connections.

These results are very encouraging. In particular those for undirected edges show a high F1-score (0.82), comparable or higher than other techniques have achieved on such complex data. There is not a clear candidate for a baseline here, though the results of this combined approach are compared to a machine learning approach and evaluated against leading techniques for other popular datasets in Section 9.6. Future work in this area will look at the development of other baselines, and explore other techniques for reporting results including the use of graph edit distance (Gao et al., 2010) and Combined Argument Similarity Score (CASS) Duthie et al. (2016b).

	Directed			**Undirected**		
	p	**r**	**F1**	**p**	**r**	**F1**
Rule based combination	0.69	0.79	0.74	0.86	0.79	0.82

Table 9.1: Rule-based combination results for identifying directed and undirected connections in the US2016G1tv corpus.

9.4 Representation

With several techniques working in combination, it is important to consider how the results will be represented in order to ensure that explainability is maintained. That is, we need to ensure that the kind of combined output shown in the top half of Figure 9.4 also includes details of where each of the added S-Nodes and edges come from.

The Argument Interchange Format representation allows us to include a locution (L-Node) for each relation (S-Node), showing that an algorithm has asserted that this relation exists. An example of this can be seen in Figure 9.5. In this diagram, the identified structure is shown on the left, with an S-Node showing support between the two proposition. An L-Node representing the assertion of the algorithm that this support relation exists is shown on the far right of the diagram, and connected to the S-Node via an Illocutionary node (YA-Node).

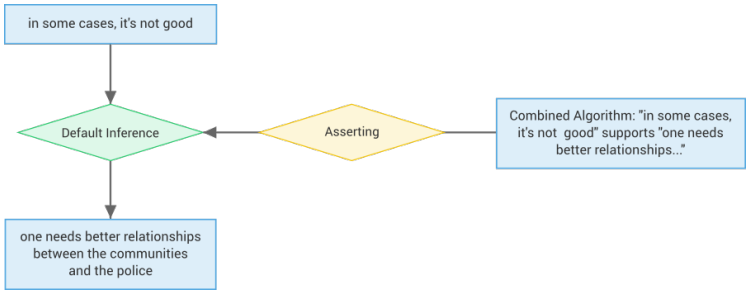


Figure 9.5: Representing the algorithm's assertion of an inference relation

In order to maintain explainability, we need to represent not just that the algorithm has asserted that this support relationship exists, but *why* the algorithm has made this assertion. One option for doing this would be to add the reason as support for the L-Node (Figure 9.6). However, this approach does not fit with the usual conception of support, that is, the presence of a discourse indicator in this case does not support the fact that the algorithm said that there is an inference relationship between the two propositions on the left.

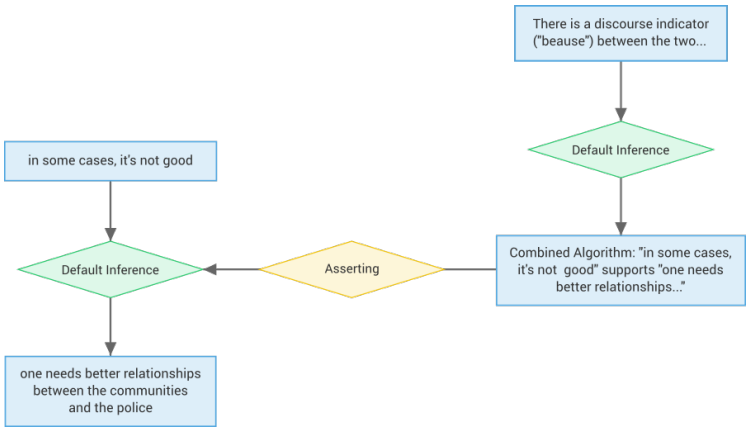


Figure 9.6: A rejected account of justifying the algorithm's assertion of an inference relation by support of the locution

In order to capture this reasoning more accurately, we turn to Searle's (Searle, 1969) account of the five rules for the use of illocutionary forces. In the case of assertion, they are as follows (here S is the speaker and H the hearer):

- The propositional content rule: what is to be expressed is any proposition p.

- First preparatory rule: S has evidence (reasons etc.) for the truth of p.
- Second preparatory rule: It is not obvious to both S and H that H knows (does not need to be reminded of, etc.) p.
- Sincerity rule: S believes p.
- Constitutive rule: Counts as an undertaking to the effect that p represents an actual state of affairs.

In this case, it is the second of these rules that we are concerned with, "S has evidence (reasons etc.) for the truth of p". The presence of a discourse indicator (or, likewise, being components of the same argumentation scheme instance, being semantically similar, etc) forms evidence for the support relationship. This representation is shown in Figure 9.7. Although the first preparatory rule is technically a component of the assertion, for the sake of simplicity, we show the reason ("There is a discourse indicator...") as supporting the YA-Node directly.

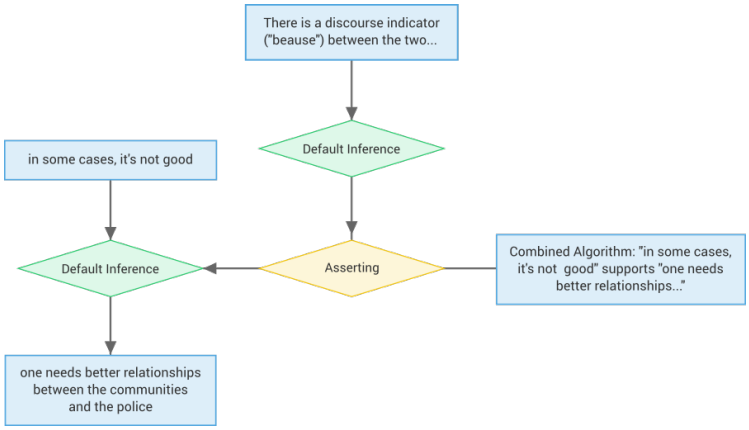


Figure 9.7: Justifying the algorithm's assertion of an inference relation by support of the assertion's first preparatory rule.

Similarly, in cases where two rules are required in order to identify a support relationship, these can be shown using the standard AIF way of representing a linked argument. An example of this can be seen in Figure 9.8, where the inference is a result of the similarity between the two propositions providing a link, and the greater centrality of "one needs relation" than "there are no good

community relations in Chicago" giving the direction of the inference between them.

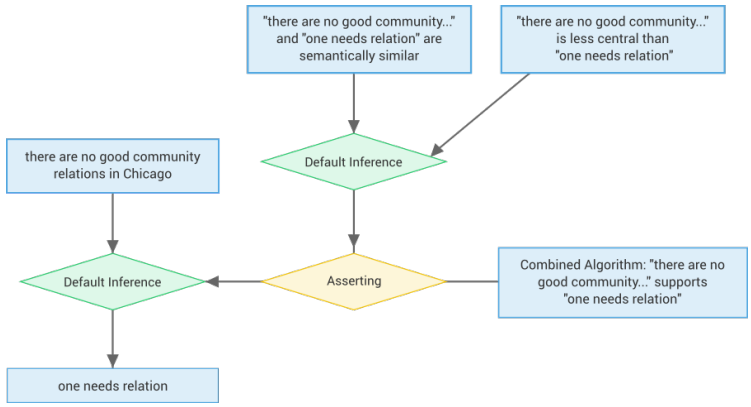


Figure 9.8: Linked support justifying the algorithm's assertion of an inference relation.

9.5 Explainability

Explainability in Artificial Intelligence is a growing concern, with, for example, 67% of business leaders surveyed in the PwC global CEO survey[1] stating that they believe "AI and automation will impact negatively on stakeholder trust levels". As the use of automated approaches for understanding human reasoning grows, it becomes increasingly important to be able to justify the information being extracted.

The term Explainable AI (XAI) was first used by Van Lent et al. (2004) to describe the ability of AI agents in a simulation game to justify their actions. However, as demand for transparency and justification in AI has grown, this term has gained a broader meaning. Adadi and Berrada (2018) define XAI as "a research field that aims to make AI systems results more understandable to humans", similarly Gunning et al. (2019) describe the purpose of an XAI system as being "to make its behavior more intelligible to humans by providing explanations".

In the field of Argument Mining, the results are the argument structure extracted from a piece of natural language text, and the behaviour of an Ar-

[1] `https://pwc.to/2pZTNuJ`

gument Mining system is to make decisions about the nature of this structure. It is exactly these decisions made about how the argument structure is extracted from the text that Explainable Argument Mining (XAM) makes more intelligible to a human user.

If a human analyst were to annotate the argument structure in the excerpt from US2016G1tv given in Figure 9.1, it would be possible to ask them questions such as "why have you added an inference relation between 'in some cases, it's not good' and 'one needs better relationships between the communities and the police'?" These are the kind of questions that XAM system must be able to answer.

Section 9.4 has shown how the reasoning behind such decisions made by the combined XAM system presented here can be captured in the Argument Interchange Format (AIF). This reasoning can also be conveyed in natural language generated from the AIF structure using simple templates. For example, a natural language version of the reasoning captured in Figure 9.8 could read as shown in Example (1), answering the question of why the system has added this inference relation. Similar explanations for the relations labelled '1' and '2' in the top half of Figure 9.4, are given in Examples (2) and (3) respectively[2].

(1) **An *RA*-node with the scheme** *"Default Inference"* **has been added between** *"there are no good community relations in Chicago"* **and** *"one needs relation"* **because** *the word vector similarity score for the two segments is over the threshold of 0.310* **AND** *"there are no good community..." is less central than "one needs relation".*

(2) **An *RA*-node with the scheme** *"Default Inference"* **has been added between** *"in some cases, it's not good"* **and** *"one needs better relationships between the communities and the police"* **because** *there is a discourse indicator ("because") between the two.*

(3) **An *RA*-node with the scheme** *"Example"* **has been added between** *"Chicago's not the only"* **and** *"in so many different places it is*

[2]In these examples text in bold is part of the template, and text in italics is filled in from the AIF representation

similar as in Chicago" **because** *"Chicago's not the only" matches the argument scheme component 'EX Premise'* **AND** *"in so many different places..." matches the argument scheme component 'EX Conclusion'.*

Such explanations offer a broad range of potential use cases, for example: highlighting unintended conclusions and suggesting how these can be fixed; helping an audience to determine whether or not they agree that the inferential relationship was intended by the speaker; and, making it easier for researchers to see where things have gone wrong when performing error analysis in comparison to gold standard annotation.

9.6 Evaluation

In this section the rule based combination approach detailed in Section 9.2 is evaluated, firstly by comparison to machine learning combination approaches, and then by testing on other widely used argument corpora and comparing to state of the art results for these.

9.6.1 Comparison to machine learning combination

The experiments in machine learning combination use a range of classifiers implemented in Scikit-learn(Pedregosa et al., 2011): Random Forest, Linear Support Vector Classification, Multinomial Naive Bayes, and Logistic Regression. In each case, the input consists of all unique pairs, a and b, of ADUs within moving window of size 10 (i.e. all pairs of ADUs that are at most 9 apart in sequential ordering). Each pair is labelled as to whether there is an inferential connection from a to b. Limiting potential connections to within this moving window is done to reduce the vast number of possible connections between distantly separated ADUs, whilst still allowing for the identification of the majority of connections ($\sim$85% of those in the gold standard data). Whilst connections are classified within the moving window described, the results presented in the next section are calculated as compared to **all** valid connections so as not to unfairly bias them in favour of this approach.

The features implemented for each classifier are as follows:

- **Discourse Indicators:** Binary features representing whether any of the indicators identified in Chapter 4 are present between a and b for sequential a and b.
- **Premise-Conclusion Topic Models:** The probabilities for each topic in the topical inference matrix, for both a and b, are added as individual features.
- **Similarity:** The similarity scores for a and b using each similarity method.
- **Graph Properties:** The centrality and divisiveness scores for both a and b.
- **Argument Schemes:** The resulting probabilities for each scheme component for both a and b.

The results from the machine learning implementations are shown in Table 9.2. Several observations can be made from these results. Firstly, the results for the rule-based combination approach are significantly better than those for any of the different machine learning classifiers. The machine learning classifiers all perform relatively similarly, suggesting that their weaker performance is not a result of any particular classifier. For the machine learning approaches, the recall is generally lower. This is due to the limitation of only identifying connections within a fixed size window, and, as such, not labelling any connections that occur beyond this distance. Whilst this reduces the recall, removing this restriction would produce a large number of false positives throughout the vast number of possible long distance connections. The precision for the machine learning methods is also lower in many cases than that for the rule-based method, suggesting that there is not enough data for an unsupervised approach to learn similar rules to the manually created ones. This situation would be expected to be even worse on many datasets which are generally smaller than US2016G1tv.

Combination Method	Directed			Undirected		
	p	**r**	**F1**	**p**	**r**	**F1**
Rule-based combination						
Rule based combination	0.69	0.79	0.74	0.86	0.79	0.82
Machine learning combination						
RandomForest	0.63	0.44	0.52	0.72	0.44	0.55
LinearSVC	0.58	0.55	0.56	0.66	0.55	0.60
MultinomialNB	0.54	0.47	0.50	0.6	0.47	0.53
LogisticRegression	0.61	0.45	0.52	0.69	0.45	0.54

Table 9.2: Rule-based and machine learning combination results for identifying directed and un-directed connections in the US2016G1tv corpus.

9.6.2 Testing on the Argument Annotated Essays Corpus

In order to test performance further, the rule-based method was applied to two widely used argumentation corpora and the results compared to existing work on these datasets. In this case the same argumentation scheme classification model was used based on training data from US2016tv (see Chapter 8). New Premise-Conclusion topic models were automatically generated using the method outlined in Chapter 5 (determining keywords from the pre-segmented corpus data, performing a web search for matching documents, etc.) The first corpus considered here is the Argument Annotated Essays Corpus (version 2) (Stab and Gurevych, 2017). The corpus contains 402 persuasive essays annotated with fine-grained argumentation structure (an example essay can be seen in Figure 9.9). On average each essay includes 18 sentences and 366 tokens, with a total across the whole corpus of 147,271 tokens and 7,116 sentences.

The annotation labels three different proposition types:

MajorClaim: the thesis statement expressing the stance of the author about the topic. In cases where this statement is present in several reformulated forms, these are all annotated as major claims

Claim: the central component of an argument either supporting or attack-

Arts and public services are both important to the community and should be invested properly

There has been wide opinion that Government should invest more money on public services rather than arts such as music and theatre. However, in my perspective, both public services and arts worth investment.

It is obvious axiomatic that indispensable public services like hospital and school worth every penny investing. Investing in hospitals helps improve people's health, also, in schools assists with raising our children's education level. Apparently if these facilities were not spent adequately, our standard of living would deteriorate.

But our standard of living also depend on another factor - spiritual life which is related closely with arts. Arts include many forms and music as well as cinema are the most typical. These two art forms not only provide the public with entertainment but also contribute significantly to the economy. The income of film and music industries produce millions of dollars each year for the Goverment, for instance K-pop and Hollywood, and these industries can not survive without goverment's financial assistance.

The long and the short of it, both arts and public services are important to the community and should be invested properly.

Figure 9.9: Essay 396 from the Argument Annotated Essays Corpus

ing the major claim

Premises: the reasons given by the author for supporting or attacking the claims.

Each claim is assigned a stance either for, or against, the MajorClaim. The connections between claims and premises are annotated as either "supports" or "attacks". The annotation for the previously mentioned example essay can be seen in Figure 9.10.

In order to run the methods presented here on this data, it was first converted to AIF format using the algorithm shown in Algorithm 1. The resulting argument structure for essay 396 imported into AIFdb is shown in Figure 9.11. The complete AIF translation of the AAEC corpus is available online at `http://corpora.aifdb.org/AAECv2`.

Argument Component Classification

Stab and Gurevych (2017) consider the classification of argument component types as multiclass classification and label each argument component as "MajorClaim", "Claim", or "Premise". The methods presented in this thesis do not classify components in this way, viewing such classifications as a result of the argument structure, rather than intrinsic properties of the text. However, once the argument structure is obtained, we can replicate these results by viewing a Premise as a node with no incoming edges, a Claim as a node that

Data: essayXXX.ann file from Argument Annotated Essays Corpus

Result: AIF translation of the supplied .ann file

begin

- **foreach** MajorClaim *in* essayXXX.ann **do**
  - **if** first MajorClaim **then**
    - Add an I-node corresponding to MajorClaim text [MC1]
  - **else**
    - Add an I-node corresponding to MajorClaim text [MCn]
    - Add an MA-node [MCn → MC1]
  - **end**
- **end**
- **foreach** Claim *in* essayXXX.ann **do**
  - Add an I-node corresponding to MajorClaim text [Cn]
- **end**
- **foreach** Premise *in* essayXXX.ann **do**
  - Add an I-node corresponding to Premise text [Pn]
- **end**
- **foreach** Stance *in* essayXXX.ann **do**
  - **if** type = For **then**
    - Add an RA-node [Cn → MC1]
  - **else if** type = Against **then**
    - Add a CA-node [Cn → MC1]
  - **end**
- **end**
- **foreach** supports *in* essayXXX.ann **do**
  - Add an RA-node [Arg1 → Arg2]
- **end**
- **foreach** attacks *in* essayXXX.ann **do**
  - Add an CA-node [Arg1 → Arg2]
- **end**

end

Algorithm 1: Conversion of AAEC annotations to AIF

T1 MajorClaim 254 300 both public services and arts worth investment
T2 MajorClaim 1181 1273 both arts and public services are important to the community and should be invested properly
T3 Claim 331 413 indispensable public services like hospital and school worth every penny investing
A1 Stance T3 For
T4 Premise 415 537 Investing in hospitals helps improve people's health, also, in schools assists with raising our children's education level
T5 Premise 550 637 if these facilities were not spent adequately, our standard of living would deteriorate
R1 supports Arg1:T4 Arg2:T3
R2 supports Arg1:T5 Arg2:T3
T6 Premise 749 821 Arts include many forms and music as well as cinema are the most typical
T7 Premise 823 938 These two art forms not only provide the public with entertainment but also contribute significantly to the economy
T8 Premise 940 1148 The income of film and music industries produce millions of dollars each year for the Goverment, for instance K-pop and Hollywood, and these industries can not survive without goverment's financial assistance
T9 Claim 643 747 our standard of living also depend on another factor - spiritual life which is related closely with arts
A2 Stance T9 For
R3 supports Arg1:T6 Arg2:T9
R4 supports Arg1:T7 Arg2:T9
R5 supports Arg1:T8 Arg2:T9

Figure 9.10: Annotation of essay 396 from the Argument Annotated Essays Corpus

has both incoming and outgoing edges, and a MajorClaim as a node which has no outgoing edges.

In order to test these methods on the AAEC corpus, the combined approach detailed in Section 9.2 was applied, and from the results the labels "Majorclaim", "Claim", and "Premise" were calculated. We then compare accuracy of these to the gold standard data, with results as shown in Table 9.3. In this table the first three columns show the combined F1, precision, and recall for classification of all argument components (MajorClaim, Claim and Premise) in AAEC. The remaining three columns then show the individual F1 results for each of MajorClaim, Claim and Premise.

Stab and Gurevych (2017) provide a majority baseline for these results based on the classification of all components as Premise (the full corpus contains 751 major claims, 1,506 claims, and 3,832 premises) along with a heuristic baseline motivated by the common structure of persuasive essays. The heuristic baseline for the argument component classification task labels the first argument component in each body paragraph as a claim, and all remaining components in body paragraphs as premise. The last argument component

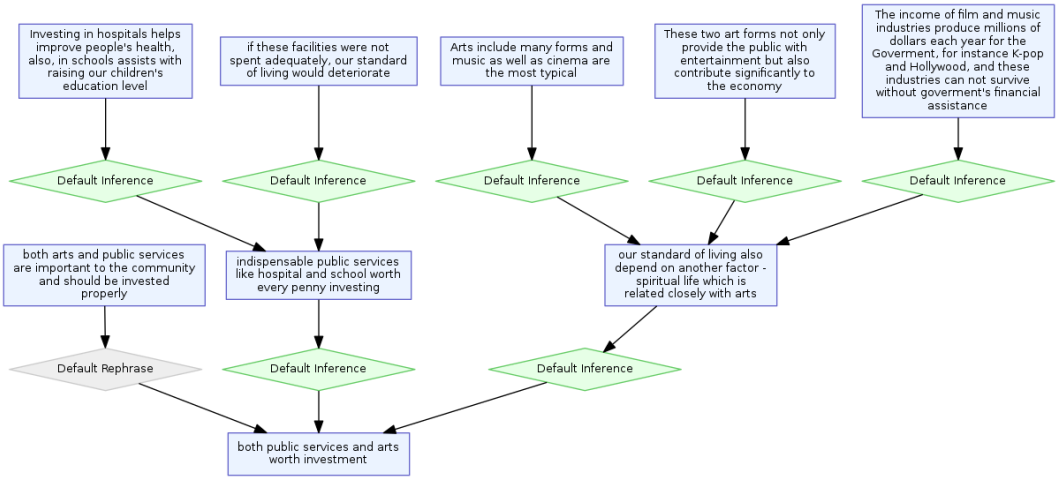


Figure 9.11: Essay 396 from the Argument Annotated Essays Corpus imported into AIFdb

Method	**F1**	**p**	**r**	**F1 MC**	**F1 Cl**	**F1 Pr**
Stab and Gurevych (2017)						
Baseline majority	0.26	0.21	0.33	0	0	0.77
Baseline heuristic	0.72	0.72	0.72	0.74	0.56	0.87
SVM all features (best classifier)	0.77	0.77	0.77	0.87	0.59	0.86
ILP-balanced (best joint model)	0.82	-	-	0.87	0.70	0.90
Rule based combination						
Rule based	0.75	0.71	0.79	0.81	0.67	0.76

Table 9.3: Overall F1, precision, and recall for argument component classification on AAEC, and individual F1 scores for MajorClaim, Claim and Premise

in the introduction and the first argument component in the conclusion are classified as major claims and all remaining argument components in the introduction and conclusion are labelled as claims. Whilst the majority baseline achieves weak results, with an overall F1-score of 0.26, the heuristic baseline is much more challenging, with an overall F1-score of 0.72. In both cases, the combined rule-based approach presented here beats these baselines (with an overall F1-score of 0.75), however the heuristic baseline slightly outperforms the combined approach in terms of precision. This result seems likely to be due to a slight tendency for the combined approach to prefer the Claim label (finding components with both edges in and out is made slightly more likely)

and indeed Claims have the lowest individual F1-score with 0.67, compared to 0.81 for MajorClaim and 0.76 for Premise. Considering the challenging nature of the baseline and the differences between this data and the original target of US2016G1tv, these are still extremely encouraging results.

Stab and Gurevych (2017) train two base SVM classifiers, one to recognise the type of argument component, and another to identify argumentative relations between argument components. SVMs were selected in this case as they have been shown to outperform several other learners in both tasks (Stab and Gurevych, 2014b). The outcomes of both classifiers were globally optimised in order to find the optimal argumentation structure using Integer Linear Programming (ILP). Table 9.3 shows the results for both the best performing SVM argument component classifier, and the best performing ILP model for this part of the task (precision and recall were not individually reported for the ILP models). Although the SVM classifiers are trained using a range of feature types (lexical, structural, contextual, syntactic, probability, indicators, discourse, and embedding), the structural features are the only ones that, when used on their own, significantly outperform the F1-score of the heuristic baseline. These structural features include whether the component is first or last in paragraph and whether the component is present in the introduction or conclusion. As with the heuristic baseline, such features are strongly tied to the argument component classification for this data, and help both the SVM and ILP results considerably. Despite not utilising this information, the combined approach presented in this chapter achieves an overall F1-score close to that for the best performing SVM (0.75 compared to 0.77 for the SVM) and remains reasonably competitive with the ILP model results (F1-score of 0.82).

Argumentative Relation Identification

The relation identification model (Stab and Gurevych, 2017) classifies ordered pairs of argument components as "linked" or "not-linked". In this analysis step, we consider both argumentative support and attack relations as "linked". Results are shown in Table 9.4.

As with the argument component classification task, Stab and Gurevych

Method	F1	p	r	F1 NO Link	F1 Linked
Stab and Gurevych (2017)					
Baseline majority	0.46	0.42	0.50	0.91	0
Baseline heuristic	0.66	0.66	0.66	0.89	0.44
SVM all features (best classifier)	0.73	0.76	0.71	0.92	0.54
ILP-balanced (best joint model)	0.76	-	-	0.92	0.60
Rule based combination					
Rule based	0.73	0.78	0.69	0.78	0.71

Table 9.4: AAEC Argumentative Relation identification results.

provide a majority baseline (NO Link) and a more challenging heuristic baseline. The heuristic baseline for the relation identification task classifies an argument component pair as linked if the target is the first component of a body paragraph, this baseline is based on the fact that 62% of all body paragraphs in the corpus start with a claim. Again, the combined approach presented here beats both of these baselines, with a combined F1-score of 0.73 compared to 0.46 for the majority baseline and 0.66 for the heuristic baseline.

Table 9.4 also shows the results for Stab and Gurevych's best performing SVM classifier and ILP model applied to the argumentative relation identification task. As with argument component classification, these results are likely helped by the inclusion of structural features (with, as previously discussed, 62% of all body paragraphs starting with a claim) which the combined approach does not utilise. However, in this case the combined results go as far as to match the SVM classifier (with both achieving an F1-score of 0.73) and fall only 0.03 behind the ILP model (F1-score = 0.76).

These results are extremely encouraging and show that, for tasks such as argumentative relation identification, the combined approach presented here can be competitive with other state of the art techniques, whilst maintaining its ability to explain the structural classifications that have been made, and work across genres without modification.

9.6.3 Testing on the Argumentative Microtext Corpus

The Argumentative Microtext Corpus (Peldszus, 2014) contains 112 short texts generated in a controlled text generation experiment. In the text generation experiment, participants were instructed to write a text on a topic chosen from a given set of trigger questions, with the instructions that each text should: be about five segments long; contain only segments that are argumentatively relevant (either the main claim of the text, supporting the main claim or another segment, or attacking the main claim or another segment); contain at least one possible objection to the claim; be written in such a way that it would be understandable without having its trigger question as a headline. The corpus contains the original texts with accompanying annotation based on Freeman's theory of argumentation structure (Freeman, 1991, 2011), that is, viewed as a hypothetical dialectical exchange between a proponent, who presents and defends his claims, and an opponent, who critically questions them. These moves can then be represented as an argument graph, with the nodes representing the propositions expressed in text segments and the edges between them representing different supporting and attacking moves. An example MicroText can be seen in Figure 9.12, with the corresponding AIF structure[3] shown in Figure 9.13.

> Yes, it's annoying and cumbersome to separate your rubbish properly all the time. Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins. But still Germany produces way too much rubbish and too many resources are lost when what actually should be separated and recycled is burnt. We Berliners should take the chance and become pioneers in waste separation!

Figure 9.12: MicroText 001 from the Argumentative Microtext Corpus

For the purposes of comparison, the results presented in (Peldszus, 2018) are used here. These results were obtained using an evidence graph model, where base classifiers were first trained to classify for the role (is the segment of the proponent or opponent role), function (does the segment present the central claim of the text, or does it support or attack another segment) and central claim level (the probability of the segment being a central claim). From

[3]The import of the MicroText corpus to AIFdb was completed as part of a student project and is not a contribution of this thesis.

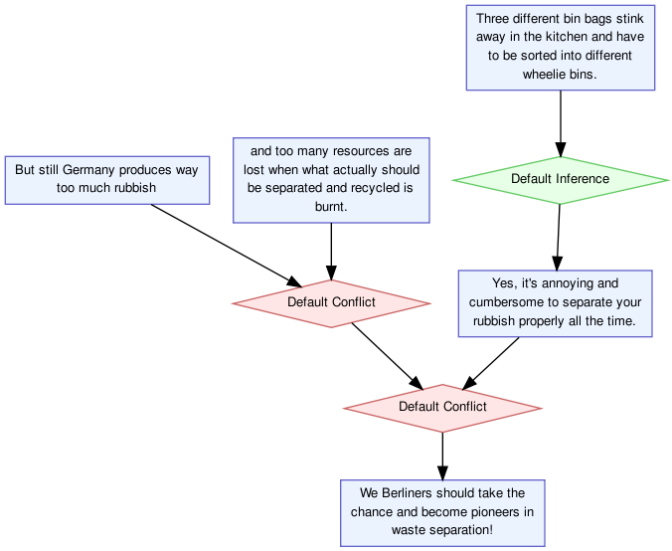


Figure 9.13: MicroText 001 from the Argumentative Microtext Corpus imported into AIFdb

these results, the evidence graph was constructed by building a fully connected multigraph over all segments with as many edges per segment-pair as there are edge types, then translating the segment-wise predictions into level-specific edge scores. Table 9.5 shows a comparison between the results of (Peldszus, 2018) and the rule based combination method presented in this chapter. As with the AAEC, the same argumentation scheme classification model was used from Chapter 8 and new Premise-Conclusion topic models were automatically generated using the method outlined in Chapter 5.

	p	**r**	**F1**
Detecting central claims			
Peldszus (2018)	0.80	0.80	0.80
Rule based combination	0.76	0.81	0.78
Identifying support relations			
Peldszus (2018)	0.76	0.82	0.79
Rule based combination	0.77	0.80	0.78

Table 9.5: Comparison of results from Peldszus (2018) against the combined rule based approach for the Argumentative Microtext Corpus

For the combined method, central claims are identified as being those with

only incoming edges and no outgoing edges. Comparing the central claim detection results of the combined approach with those of (Peldszus, 2018), we can see that the F1-score is comparable (0.80 for Peldszus and 0.78 for the combined approach). We also see that whilst the recall is slightly better (0.80 and 0.81 respectively for the two approaches) the precision is 4 points lower (0.76 versus 0.80). These results suggest that the combined approach is leaning slightly in the direction of classifying segments as central claims when they are not. Further analysis suggests that this is occurring in cases where a segment which should be labelled as a premise with one outgoing edge is instead being labelled as having an incoming edge. This issue is most likely due to the small number of nodes in each Microtext example making it harder to determine centrality, and therefore edge direction. Comparing these results to those in Section 9.6.2 we can also see that detection of MajorClaims in the AAEC slightly outperforms the detection of Central Claims in the Microtext corpus, again suggesting that longer texts with more data are easier for the methods presented here to classify.

A similar situation is also seen for the results of the support relation identification task (Table 9.5). Again precision is lower than recall, though for this task, these results reflect those of (Peldszus, 2018). The overall F1-Score for the combined approach is extremely competitive (0.78 compared to Peldszus' 0.79). Despite the challenges posed for the combined approach in dealing with small scale examples, these results again show that this approach can produce results comparable to those obtained by techniques specifically developed for these datasets, whilst also having the advantage of providing explainable reasons behind the structure identified.

9.7 Conclusion

In this chapter, we have looked at the combination of the individual explainable argument mining techniques that have been presented thus far in this thesis. After all, as each of these techniques draws inspiration from the ways in which humans understand the structure of an argument, we can also consider that

in human understanding, such clues to the processing of a complex argument work in unison.

A rule-based method of combination has been presented, which has the advantage of maintaining the explainability inherent in each individual approach. For example, the combined results allow us to say that the system believes there is an inferential relationship between x and y because they form part of an argument scheme instance, and between y and z because there is a discourse indicator (e.g. "because") between them in the original text. A representation of these results, compliant with the Argument Interchange Format (Chesñevar et al., 2006) and Inference Anchoring Theory (Budzynska et al., 2014) has been proposed. This representation shows how the reasons for the decisions made can be viewed as supporting Searle's (Searle, 1969) first preparatory rule for assertion (the speaker has evidence (reasons etc.) for the truth of the proposition being asserted).

The rule based combination method has been compared to four different machine learning based methods of combination, and shown to outperform these in every case. This has shown that not only does such a rule-based approach maintain explainability, but does not lose out in performance compared to alternative combination approaches. Finally, the rule based combination method has been evaluated against two widely used argumentation corpora (the Argument Annotated Essays Corpus, and the Argumentative Microtext corpus) and the results compared to existing work on these datasets. This comparison has shown that the combined approach can produce comparable results to state of the art techniques developed specifically for use on this data.

Chapter 10

Applications of Argument Mining

In this chapter, a number of potential downstream applications of argument mining are presented. These range from applications which rely directly on argument mining algorithms to provide their functionality (such as the Evidence Toolkit covered in Section 10.1) to software for visualising and analysing arguments once the argumentative structure has been successfully mined (Section 10.4). It should be noted that this chapter does not aim to provide a comprehensive overview of such applications and software, but to cover those that have been developed by the author and that are closely related to the work presented in the rest of this thesis.

10.1 The Evidence Toolkit

The goal of The Evidence Toolkit[1] (Visser et al., 2020b) is to encourage and develop critical thinking skills, in particular as related to print and online media. The software was developed as part of "BBC School Report 2018"[2] to provide 16 to 18-year-old secondary school pupils with tools to help them understand and critique the argumentative structure present in news articles. To achieve this, the user is first walked through a selection of five pre-annotated

[1] `https://www.bbc.co.uk/taster/pilots/evidence-toolkit-moral-maze`

[2] `https://www.bbc.co.uk/academy/en/articles/art20180313125234328`

articles, where they are asked to:

1. Identify the main claim presented by the author
2. Identify reasons given by the author for this claim
3. Select the type of reason, *Fact* or *Opinion*, with sub-types of *Statistical*, *Example*, or *Other* for factual arguments, and *Expert*, *Popular* or *Personal* for opinion based arguments (see Figure 10.1)
4. Judge how well the provided reason stands up to critical questions associated with the selected evidence type
5. Identify any stated objections to the main claim, which show the author thinking about the issue from other perspectives

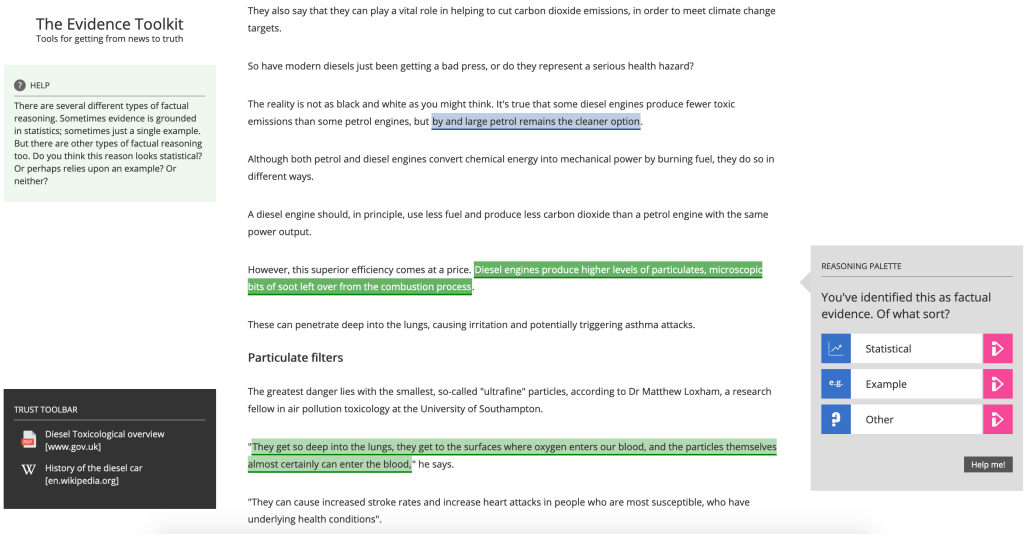


Figure 10.1: Selecting the type of an identified supporting reason in The Evidence Toolkit

Having progressed through the pre-annotated material, the user is then invited to use the *Pick Your Own* feature to carry out the same analysis on any article of their choosing from the BBC News website. For these unseen articles, there is no human annotation available, and, as such, argument mining techniques are employed to provide the user with suggestions about the main claim and associated reasons. The Pick Your Own feature combines several

argument mining techniques that have been shown to produce accurate results in previous work. Firstly, the main claim is identified by determining the centrality of each sentence in the article, that is, how semantically similar each sentence is to all of the other sentences. This method has been shown to provide a reliable indication of which claim is most central within the argumentative structure of a piece of text (see: Chapter 7; Lawrence and Reed (2017b)). Once the main claim has been confirmed by the user, or they have selected an alternative claim which they think is more central to the article, the supporting reasons for this claim, and any potential objections, are identified. To do this, all of the other sentences are ranked by their semantic similarity to the main claim, and then checked for indicators of support (e.g., "because") or conflict (e.g., "however") to determine their possible argumentative relations (see: Chapter 4; Chapter 6; Lawrence et al. (2017a)).

The Evidence Toolkit was distributed to over 3,000 educational institutions in the United Kingdom, making it the largest-scale deployment of argument mining technology available to the general public to date. Most encouragingly, 88% of users surveyed said it "changed their perception of the BBC to the positive".

10.2 BBC Moral Maze: Test Your Argument

A direct predecessor of The Evidence Toolit, *BBC Moral Maze: Test Your Argument* (Lawrence et al., 2018) is also aimed at developing critical thinking skills, though following a different approach. Rather than guiding users through the critical appraisal of news articles, Test Your Argument challenges users with a number of argumentation puzzles. The challenges help develop an understanding of the core principles of strengthening and critiquing arguments.

Test Your Argument comprises: a backend, which stores argument data, processes user selections and provides feedback and scoring on their choices; and a frontend, developed using standard web technologies (HTML5, CSS and Javascript) to ensure a consistent and visually appealing experience across a range of platforms (Figure 10.2).

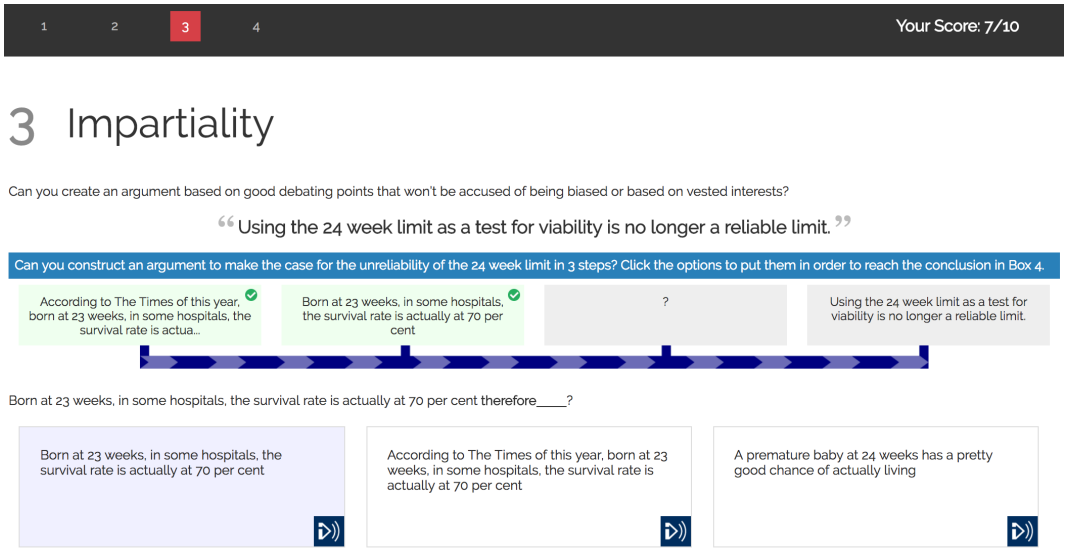


Figure 10.2: Moral Maze: Test Your Argument section 3, Impartiality

The first section presented to the user, **Strengthen**, focuses on the ways in which an argument can be strengthened and defended against attacks. The user is presented with a central statement from the debate and asked to choose, from a list three further propositions, which one best supports the statement, which one is pre-empting a counterargument, and which one attacks the opposing view.

In the second section, **Critique**, a central statement from the opposing side of the debate is given and the user is asked to consider the different types of evidence that could support this and to consider which of these might be most easily criticised. The user is asked to identify which supporting proposition is a factual statement, which is an opinion, and which is based on personal experience.

The third section, **Impartiality**, encourages considering the reasoning on both sides of an issue. The user is asked to create a chain of reasoning supporting first one side of the debate and then the other. In each case they are given three supporting statements that they have to put in the correct order to support the conclusion (see Figure 10.2).

Within each section, the user is provided with direct links to where the text appears in the Moral Maze audio on the BBC iPlayer platform. Feedback is also given for each decision that they make, with correct decisions highlighted

in green and mistakes in red, as well as a running score showing how they are progressing. At the end of the three sections, the user is able to give their own view on the issue and is provided with an aggregate score and the opportunity to share this on social media.

Since its launch in December 2017, Test Your Argument has had over 10,000 visitors, and, of those visitors that provided feedback, 80% said "Yes, the BBC should do more like this". Whilst the data used in the pilot deployment of Test Your Argument comes from a manually annotated special edition of the BBC Radio 4 programme, the *Moral Maze*[3] on the morality of abortion, future work would look to expand this scope, using automatically mined argumentative structures and allowing the user a free selection of topics over which to use the software.

10.3 Arvina & Polemicist

The web-based discussion software *Arvina* (Lawrence et al., 2012a) allows participants to debate a range of topics in real-time in a way that is structured but at the same time unobtrusive. Arvina uses dialogue protocols written using the Dialogue Game Description Language (DGDL) (Bex et al., 2014a) to structure the discussion between participants. Such protocols determine which types of moves can be made (e.g. questioning, claiming, etc.), when these moves can be made (e.g. a dialogue starts with a claim; question moves can only made in the turn directly following a claim; etc.), and describe how each move updates the argument structure of the discussion taking place.

Arvina can support multiple human users interacting in the same dialogue, as well as incorporating software agents representing (the arguments of) specific authors who have their opinions stored in AIFdb (Lawrence et al., 2012b). So, for example, say that Wilma has constructed a complex, multi-layered argument using the OVA argument analysis tool (Janier et al., 2014), concerning the use of nuclear weapons. An agent representing Wilma can then be added to an Arvina discussion and questioned about these opinions, with the agent

[3] `http://www.bbc.co.uk/programmes/b006qk11`

answering by giving Wilma's pre-annotated opinions.

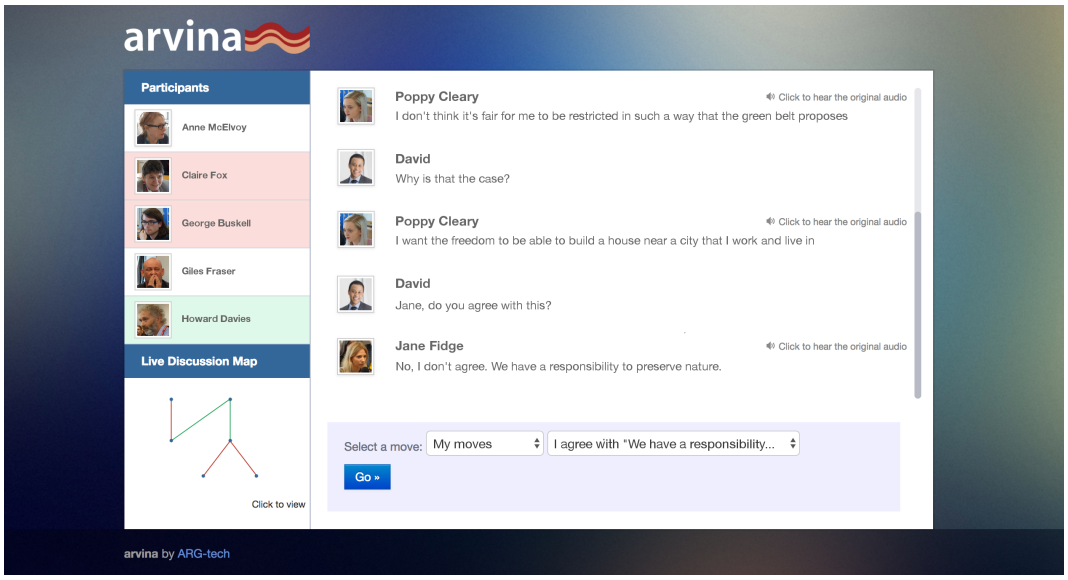


Figure 10.3: The Arvina user interface

Figure 10.3 shows the Arvina debate interface. The top left corner offers a list of participants with red or green highlighting to show their stance on the current point (either calculated from an agent's argument graph knowledge base, or provided directly in the case of a human user). Below this, there is a live discussion map showing the structure of the debate so far. This structure is also saved to AIFdb, allowing future users to interact with any points made in the current debate. On the right hand side, there is a transcription of the debate, and below this, a selection panel where the user can choose from their list of available moves at each point, and input their own opinions as the selected moves allow.

The DGDL dialogue games available in Arvina have been developed to capture a range of structured conversations, for example, to facilitate the generation of mathematical proofs (Pease et al., 2014) or allow for debate of moral issues. The *Polemicist*[4] application (see Figure 10.4), in particular, offers a custom version of Arvina, giving users the opportunity to interact with agents representing the panellists and witnesses from the BBC Moral Maze radio programme. Polemicist uses a fixed DGDL protocol, allowing the user take on the role of the moderator of the debate: selecting topics, controlling the flow of the

[4] `http://polemici.st`

dialogue, and thus exploring all the angles of the rich argumentative content. Playing the role of moderator lets the user rearrange the arguments and create wholly novel virtual discussions between the contributions of participants that did not directly engage in the original debate, while still reflecting their stated opinions.

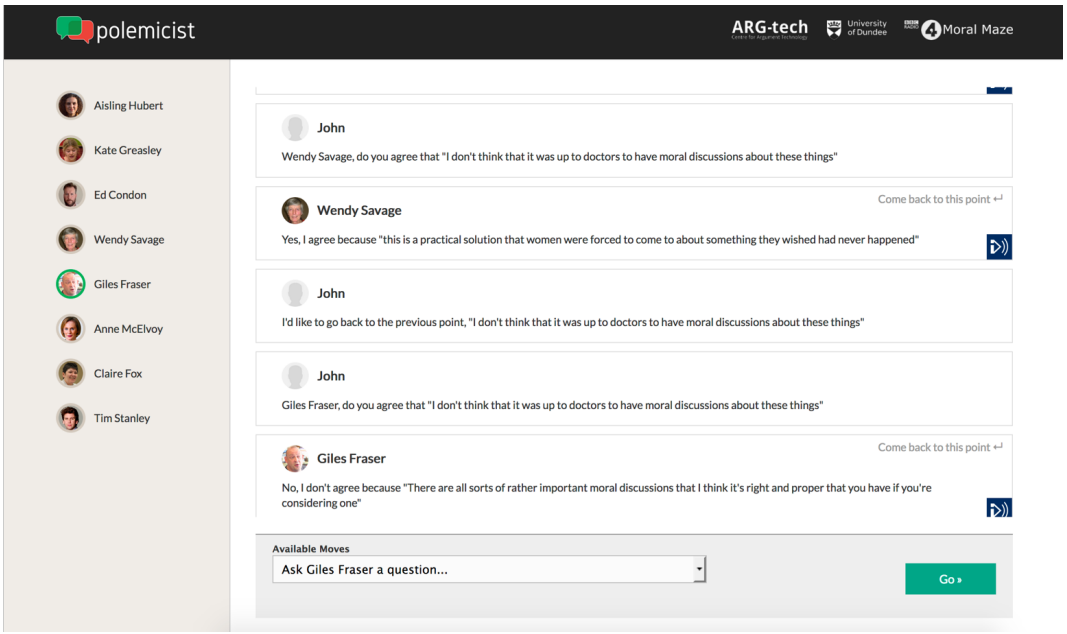


Figure 10.4: The Polemicist user interface

Whilst Arvina and Polemicist currently rely on pre-annotated material from AIFdb to provide the responses for agents in a dialogue, they represent a valuable use case for automatically mined arguments. If the argumentative structure in a radio transcription can be extracted by argument mining, conversations in Polemicist could take place as the radio programme is being transmitted. Similarly, discussions aimed at creating new mathematical proofs in Arvina could feature counter-examples extracted from online mathematical publications. Combining these dialogue interfaces with a robust argument mining platform would enable users to discuss any issue of their choosing with any person whose opinions on that topic have been previously recorded.

10.4 Argument Analytics

Argument Analytics (Lawrence et al., 2016) provides a suite of techniques for analysing and visualising features of Argument Interchange Format data, be that pre-annotated data from AIFdb, or the output of argument mining software. Argument Analytics components range from the detailed statistics required for discourse analysis or argument mining, to infographic-style representations, offering insights in a way that is accessible to a general audience. The extensible set of modules currently comprises: simple statistical data, which provides both an overview of the argument structure and frequencies of patterns such as argumentation schemes; dialogical data highlighting the behaviour of participants of the dialogue; and real-time data allowing for the graphical representation of an argument structure developing over time. Together these analytics open an avenue to giving feedback on live debates, producing summaries of deliberative democracy, mapping citizen science, and more.

10.4.1 Simple Statistics

The simple statistics modules allows an analyst to quickly make sense of a large amount of annotated argument data. Although these calculations are straightforward and relatively easy to automate, they nevertheless provide interesting insights into the data. The overview page (Figure 10.5 shows a range of statistics, offering a rapidly digested summary of the overall argumentative structure. The number of Information nodes provides an indication of the overall size of the analysis. The average number of words per Information Node illustrates the complexity of the ideas presented, and how succinctly they are expressed. The numbers of inference (RA) and conflict (CA) nodes give a suggestion as to the nature of the dialogue, which can be further explored by expanding the list of scheme instances present for each node type.

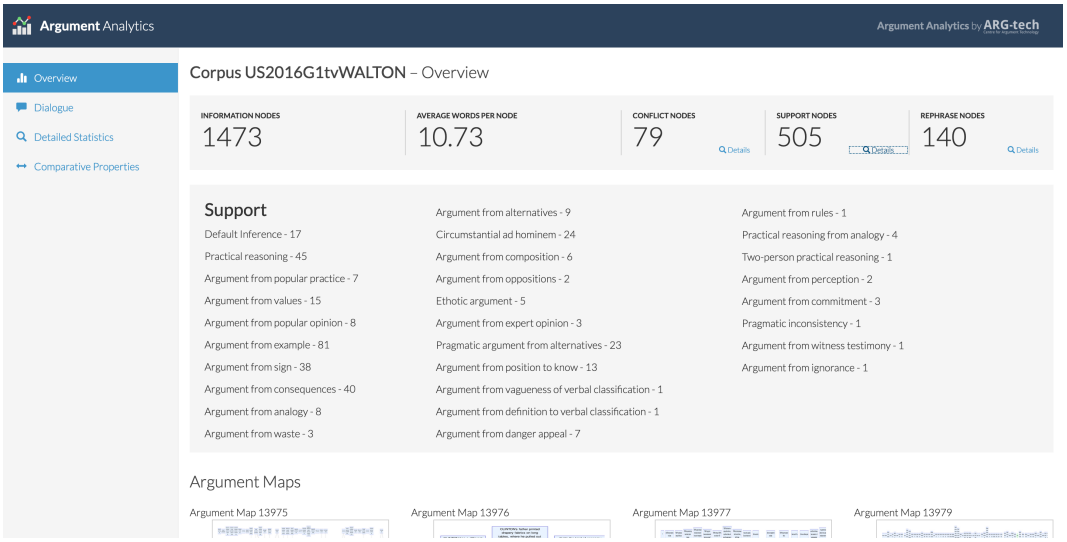


Figure 10.5: Simple statistics on the Argument Analytics Overview page

10.4.2 Dialogically Oriented Statistics

For those argument analyses where there is a dialogue taking place between multiple participants, a range of dialogically oriented analytics modules are able to provide insights into the dynamics of the discourse, and make these complex interactions accessible to a general audience. Dialogically oriented statistics currently available in the Argument Analytics suite include:

Participation: For each participant, the number of locutions they have made is counted and represented in a bar chart. This provides an easy way of identifying which participants were most, and least, dominant within a dialogue. Figure 10.6 (left) shows the participation of participants in a BBC Moral Maze radio programme.

Stimulating: A point of debate is stimulating if it receives responses, either to agree or disagree. From the analysed argument structure, we count the number of locutions which each participant has made that have at least one response, and those which have been ignored by the other participants. Figure 10.6 (right) shows the stimulating scores for each participant in a BBC Moral Maze radio programme.

Interactions: Shown as a chord diagram representing the interaction between participants (see Figure 10.7). A chord diagram is a graphical method of displaying the inter-relationships between data in a matrix. The data is ar-

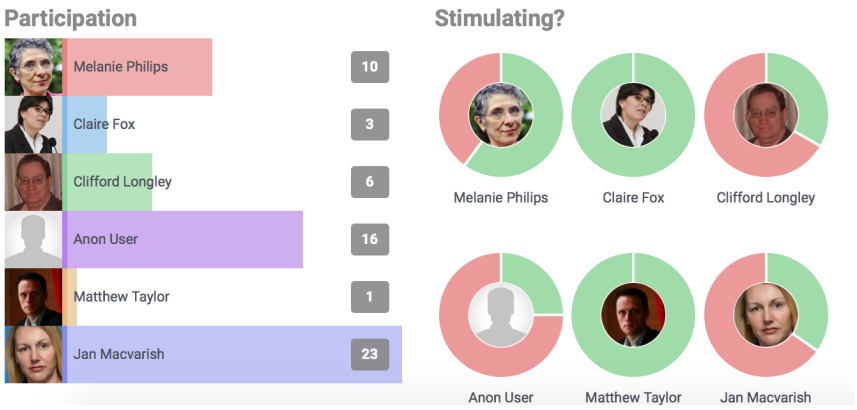


Figure 10.6: Graphical representations of the relative involvement of each participant in a dialogue, and how stimulating the points made by each participant are.

ranged radially around a circle with the relationships between the points drawn as arcs connecting the data together. In this case, the arcs represent interaction between participants, with the width of the arc at each end representing the number of locutions made by that participant to which the connected participant has responded.

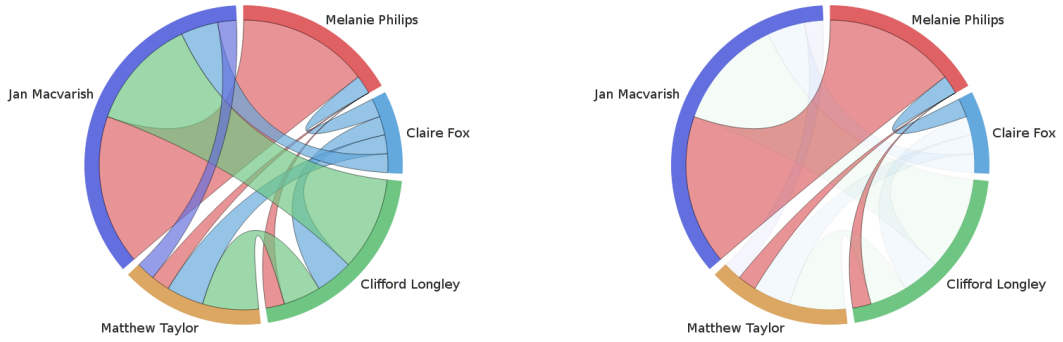


Figure 10.7: Interactions in a BBC Moral Maze episode represented as a chord diagram.

Turn Structure: Using the timestamping of locutions provided by AIFdb, a graphical representation of the turn structure in a dialogue is created. This visualisation provides a quick overview of when each participant has been most active, suggesting details of any pre-defined turn-taking rules. The example shown in Figure 10.8 reflects the turn structure in a *Moral Maze* episode. As the episode begins, each of the four regular panellists speak briefly about the topic being discussed. A guest witness is then introduced, and, after providing

their own views on the topic, are then questioned by first one of the panellists and then by a second.

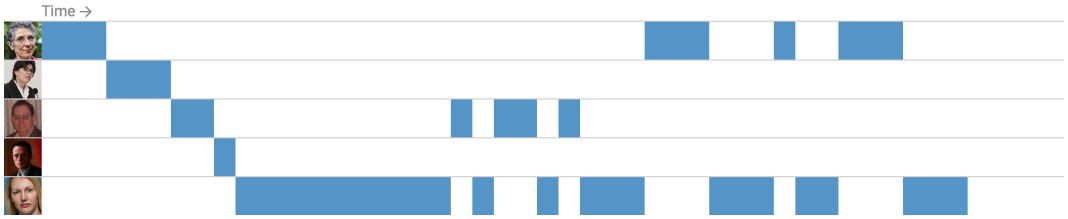


Figure 10.8: Graphical representation of the turn structure in a dialogue

10.4.3 Real-time Statistics

Many of the modules used in Argument Analytics have the ability to not only display data on a fixed, pre-analysed argument structure, but to update in real-time as the structure evolves. This functionality has been used, for example, in a tool developed for the *Built Environment for Social inclusion through the Digital Economy (BESiDE)* project[5], to facilitate round table discussions between architects working on the design of care environments, and the various stakeholders involved in the design process.

As the discussion is taking place, the audio is recorded and an analyst uses a custom-designed interface to segment the dialogue when either the topic or the speaker changes. A simple dialogue protocol is used, allowing participants to make moves of various types (e.g. asking questions, agreeing with another participant, and offering their own opinion), and relating to a set of pre-defined topics relevant to the design project.

Throughout the discussion, the dialogue overview shown in Figure 10.9 is displayed for all participants to see. This overview includes a transcript of the dialogue on the right hand side, and analytics modules displaying how much each participant has spoken, and which topics have been discussed on the left. Observing these tools in use, it is interesting to see that they serve not only an informative function, but actually impact the dynamics of the dialogue. When a participant can see that they are talking more than everyone else, they tend to let others speak more. When someone hasn't spoken yet, the other

[5] `http://beside.ac.uk/`

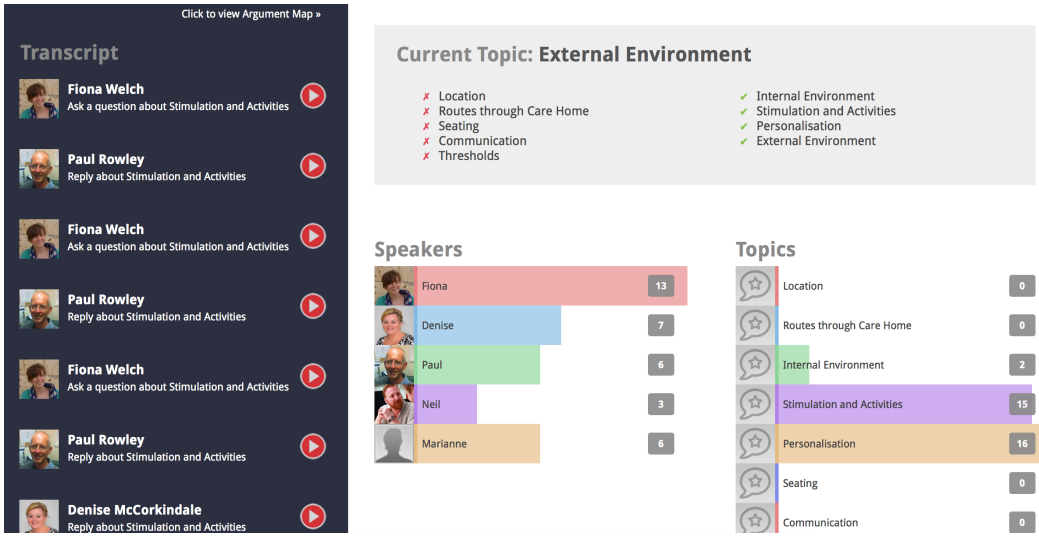


Figure 10.9: Real-time Argument Analytics highlighting the involvement of individual participants and the topics discussed.

participants notice this, and make an effort to direct questions at them. And, when one topic has been less explored than the others, there is a noticeable shift towards that area in both the questions asked and the points raised.

This ability for the argumentative and dialogical structure to not only represent the outcome of a discussion, but to inform the participants and help ensure that all areas are fully explored has wide ranging potential applications. The current limitation to providing this kind of interface more widely is the ability to perform real-time analysis, but as argument mining techniques improve, these could hopefully be combined with current state-of the-art speech to text algorithms to provide analytics such as those shown here, as a dialogue is taking place.

Chapter 11

Conclusion

This final chapter presents a summary of the contributions detailed in the preceding chapters. This is then followed by an exploration of potential avenues for future development, and, finally, by some brief concluding remarks on the work presented here as a whole.

11.1 Contributions

The work presented in this thesis has contributed the following advances in argument mining and related fields: the introduction of *Explainable Argument Mining* (XAM); the development of a range of XAM techniques, which either extend and enrich existing approaches (e.g. discourse indicators, and argumentation schemes), or open up completely new directions (e.g. premise-conclusion topic models, and graph properties); the development of a software framework for analysing, storing and working with argument data; and, the development of applications which can make use of automatically mined argument structures.

11.1.1 Explainable Argument Mining (XAM)

This thesis introduces the concept of *Explainable Argument Mining* (XAM). Where the majority of argument mining approaches to date have started from a computational linguistic perspective, applying CL techniques to identify specific facets of the argumentative structure, XAM starts by looking at the rich

heritage of philosophical research in the analysis and understanding of argumentation, and by drawing inspiration from the ways in which humans understand the structure of an argument.

Taking cues ranging from the most straightforward (such as there being a discourse indicator between two propositions), through to more complex interactions (such as: two propositions having high semantic similarity, with one being more central to the discussion than the other; or, two propositions being part of the same argumentation scheme), XAM has been shown to produce comparable results to existing approaches whilst also explaining where these results come. With future work to combine the resulting explanation structure from Chapter 9 with templated text generation, these results could be presented in a way that would be intuitively understandable to a general audience, hiding the specific details of thresholds etc. and instead saying "**A** is identified as a reason for **B** because they are similar and **B** is a more central point in the dialogue".

In Chapter 9, a rule based method for combining multiple XAM techniques was developed and evaluated against the US2016G1tv corpus, and two widely used third-party argumentation corpora (the Argument Annotated Essays Corpus, and the Argumentative Microtext corpus). These three datasets represent a broad spectrum of different domains and data types, with variance in: length, with each of the 112 Microtexts being 50-100 words, the 402 essays in AAEC averaging 366 words per essay, and US2016G1tv being one 17,190 word debate transcript; participants, with AAEC being monological, Microtexts being constructed dialogues, and US2016G1tv being fully dialogical; and, genre, AAEC being persuasive essays, Microtexts being short and clearly structured, and US2016G1tv being an hour long debate transcript.

Despite this, the combined XAM approach presented here works well across all three datasets, producing comparable results to state of the art techniques developed specifically for use on these AAEC and the Microtext corpus. These results are particularly encouraging as the combined XAM approach is not tuned for features of any of these datasets, while, for example, the classifiers of Stab and Gurevych (2017) are able to take advantage of structural features

of the specific dataset they are working with. For example, whether the component is first or last in a paragraph and whether the component is present in the introduction or conclusion, both features strongly tied to the argument component classification for this data.

As the use of automated approaches for understanding human reasoning grows, it becomes increasingly important to be able to justify the information being extracted. For example, an algorithm extracting information from a debate such as that in the US2016G1tv corpus, needs to not only be able to say "Clinton said we should do X because Y", but explain exactly which features of the text make this inference clear. Doing so allows a general audience to determine whether they agree that the inferential relationship was intended by the speaker, and researchers to perform better error analysis and adjust algorithms accordingly.

XAM has already been adopted as a key component of the BBC Evidence Toolkit project[1], and offers potential for further tools that help people to both understand the arguments that they hear, and construct better arguments themselves (see Chapter 10). For example the approaches presented here could be used to highlight possibly unintended conclusions (offering an explanation of how to clarify this), or to highlight conclusions where the support is not clearly stated.

11.1.2 Analysis of Discourse Indicators as an argument mining technique

Discourse indicators have been previously used as a component of argument mining techniques, however, there has been little previous study of how well indicators perform on their own, how frequently they occur in real-world text, and how well different individual indicators map to specific argumentative relations. In Chapter 4, such properties of discourse indicators were investigated more closely, covering indicators from existing literature, as well as those identified from annotated argument data. The results from this study showed that in the US2016G1tv corpus over 85% of inference relations were

[1] `https://www.bbc.co.uk/taster/pilots/evidence-toolkit-moral-maze`

not marked by any discourse indicator. Furthermore, several indicators which are commonly mentioned in the literature as being useful for identifying argumentative structure rarely occur in this dataset: for example, "therefore" only had one occurrence within the entire debate transcript.

Of those indicators which appeared more frequently, most provided little information. For example, whilst there were 30 instances of the indicator "so" occurring between adjacent spans, only 37.5% of these instances were between spans where a support relation existed.

The one exception to this was found to be the indicator "because". This indicator appeared between adjacent spans 71 times and, of these, 87.3% were connected by a support relationship. Whilst this is a promising result, and suggests that, in those cases where "because" occurs, it can tell us with high accuracy the type of connection, using this method on its own would still leave approximately 80% of support relations (as well as all conflict relations) unidentified.

11.1.3 Mining Argumentation Scheme structures

Although previous attempts have been made to automatically identify instances of argumentation schemes, most notably Feng and Hirst (2011), these have relied on some prior analysis of the text having taken place. In the case of Feng & Hirst for example, the starting point is text which has previously been annotated for its argumentative structure. Another possible approach is suggested in Cabrio et al. (2013), where the connection between argumentation schemes and discourse relations is highlighted, however, this requires these discourse relations to be accurately identified before scheme instances can be determined.

The approach presented in Chapter 8 overcomes this requirement by looking at the features of each component part of a scheme, enabling the identification of these scheme components in completely unanalysed text. Once these scheme components have been identified, we are able to group them together into specific scheme instances and thus obtain a full annotation of scheme instances. This work has shown that by considering the features of the indi-

vidual types of premise and conclusion that comprise a scheme, it is possible to reliably classify these scheme components. The results for this approach (F-scores between 0.78 and 0.91) were comparable results to those of Feng & Hirst (accuracies of between 0.63 and 0.91) on the same dataset, whilst removing the requirement for pre-annotated text, and allowing us to determine not just that a particular scheme was being used, but to correctly assign each proposition to its schematic role.

11.1.4 Premise-Conclusion Topic Models

Premise-Conclusion Topic Models (Chapter 5) offer a new cross-domain, unsupervised approach to argument mining. The intuition underlying these models is that there are rich and predictable thematic and lexical regularities present in the expression of human reasoning, and that these regularities can be identified and help to extract the structure of such reasoning.

By first performing an online search for material on the target topic, and then using a discourse indicator with proven high precision (e.g. "because") to identify inferential pairs contained within this data, we automatically build a dataset of common arguments within the target domain.

Generating topic models from these inferential pairs, allows us to extract patterns of specific topics being regularly used to support other topics. For example, in debates concerning abortion, arguments are carefully marshalled on both sides, with religious themes more typically appearing on one side, and feminist philosophy themes more typically on the other. For a debate on the construction of a new road, we may expect to find environmental issues on one side and economic concerns on the other.

The Premise-Conclusion Topic Model approach was shown to be effective in tackling the challenging high-level pragmatic task of identifying both connectedness and directionality between argumentative discourse units. This outcome represents strong performance for this level of task, giving results comparable to those of (Palau and Moens, 2009). Furthermore, where existing approaches are often constrained in their generality by a lack of appropriately annotated, domain-specific, data, the same requirement does not apply in this

case.

These results show a clear link between the words used to express an argument and its underlying structure, and strongly support the intuition that understanding the structure of an argument can require not only consideration of the text itself, but contextual knowledge and understanding of the broader issues.

11.1.5 Graph Properties

As with Premise-Conclusion Topic Models, the work presented in Chapter 7 on Graph Properties also offers an entirely new approach for argument mining. In this case, considering the insights that can be gained by looking at large scale argument networks as a whole. In particular the properties of *Centrality*, which can be viewed as how important an issue is to the argument as a whole, and *Divisiveness*, how much an issue splits opinion.

Centrality of propositions was calculated by determining the lexical and semantic similarities between all proposition pairs and then computing eigenvector centrality on a graph with edge weights corresponding to these similarity scores.

Divisiveness was calculated for each proposition pair by first determining the difference in positive/negative polarity between the elements in the pair and then multiplying this by their similarity score to determine a conflict score: propositions which are talking about the same thing, but have different polarity are likely to be in conflict; propositions which have the same polarity or are talking about different topics are not likely to be in conflict. From here, these values were then multiplied by the centrality score for the pair to give a divisiveness score: a divisive issue being one where there is conflict between two central issues.

Chapter 7 also showed how these measures can be combined with existing argument mining approaches to give improved results, choosing the mined structure where the centrality and divisiveness as calculated from the argument graph most closely matches with that calculated directly from the text.

11.1.6 Study of similarity techniques for Argument Mining

Chapter 6 considered how various measures of the similarity between propositions map to their argumentative relationship. This study covered a broad range of lexical (ROUGE-1, ROUGE-2, ROUGE-L, and Levenshtein Edit Distance), semantic (WordNet, word vectors, and document vectors) and topical (Latent Dirichlet Allocation) similarity measures.

Experiments were performed to determine the connection between each of these similarity measures and argumentative relations. Firstly, the similarity scores for all pairs of connected propositions in the US2016G1tv corpus were calculated, and the average of these compared to the average similarity score for all non-connected propositions. The results show a highly significant ($p < 0.001$) difference between the similarities of related and un-related propositions, as calculated by the majority of measures. The exceptions to this were the results obtained using Document Vectors (which were, nonetheless, still significant, $p < 0.05$), and those for WordNet and LDA, which showed no significant difference between the average values.

This parallel between similarity and argumentative relation was shown to be even stronger for adjacent propositions, with a significant difference (though slightly reduced in the case of ROUGE-1) between related and un-related propositions for all of the techniques. Whilst semantic similarity was shown to correlate strongly with long distance argumentative relations, reflecting the likelihood that, when a speaker is referring back to a previous point in the dialogue, they may paraphrase the original point retaining its semantic meaning, but using different words to when it was originally uttered.

Finally, applying these similarity measures directly to mining arguments was shown to work best when **any** of the measures was above a given threshold. This result shows that similarity is not just a case of one type or another mapping best to argumentative relations, but that the type of similarity can vary from one situation to another, with some argumentatively related propositions sharing a large number of words in common, others being semantically similar without sharing very many common words, and others being topically

similar but not fulfilling either of the other criteria.

11.1.7 Minor Contributions

As part of the framework in which the work in this thesis was carried out, a set of tools were developed by the author for the annotation, storage and collection of argument data. These include:

- **Online Visualisation of Argument (OVA)** (Janier et al., 2014) The most widely used tool for argument analysis, with over 2,000 users in 38 countries having produced ∼75,000 analyses since 2015.
- **AIFdb** (Lawrence et al., 2012b) An openly accessible database of argument, containing over 18,000 Argument Interchange Format (AIF) (Chesñevar et al., 2006) argument maps, with over 2.1m words and 200,000 claims in fourteen different languages[2]
- **AIFdb Corpora** (Lawrence and Reed, 2014) Collecting over 8,000 of the 14,000 analyses contained in AIFdb into a range of corpora which are publicly available in perpetuity at fixed permalinks.

Similarly, many of the applications developed by the author, and detailed in Chapter 10 have found widespread usage. In particular:

- **Argument Analytics** (Lawrence et al., 2016) Piloted in association with BBC Radio 4 Moral Maze[3]
- **BBC Moral Maze: Test Your Argument** (Lawrence et al., 2018) A tool offering users the opportunity to hone their critical thinking skills using arguments from topical discussions. Test Your Argument has over 5,000 unique users, with 80% of users surveyed saying "Yes, the BBC should do more like this".

[2]Amharic, Chinese, Dutch, English, French, German, Hindi, Italian, Japanese, Polish, Portuguese, Russian, Spanish and Ukrainian

[3]`https://www.bbc.co.uk/programmes/p05jp46h/p05jp46x`

- **The Evidence Toolkit** Deployed to over 3,000 educational institutions in the United Kingdom, making it the largest-scale deployment of argument mining technology available to the general public to date. 88% of users surveyed said it "changed their perception of the BBC to the positive".

Finally, the Literature Review in Chapter 2 forms part of a comprehensive survey of argument mining, published as (Lawrence and Reed, 2020).

11.2 Future Work

The research undertaken in this thesis highlights several potential avenues for future work. These include additional XAM techniques, and further development of the techniques presented here.

11.2.1 Rhetorical Figures

Classical approaches to argumentation theory split the topic into several areas, including Logic and Rhetoric. Whilst most approaches to argument mining have focused on issues that would, by this classification, fall under the topic of Logic, almost no work has tackled Rhetoric.

Depending on how they are classified, there are somewhere in the range of 700 rhetorical figures, or figures of speech[4], ranging from those that are familiar and in some cases well-studied in computational linguistics (such as metaphor and metonymy) to those that are obscure, complex or peculiarly specific (such as anemographia, the creation of an illusion of reality through description of the wind; and antiprosopopoeia, the representation of persons as inanimate objects).

Preliminary work has been carried out exploring the usage of rhetorical figures as a tool in (explainable) argument mining (Lawrence et al., 2017c). The goal of this work has been to highlight the value and importance of the area of rhetoric in argument mining. This work has shown that a diverse set

[4] `http://rhetoric.byu.edu/`

of rhetorical figures can be identified – indeed, the very definition of many figures serves as the algorithm for their identification. The second step is to show that the consideration of rhetorical figures allows the formulation of new and intriguing hypotheses: Does polyptoton (repeating a word, but in a different form) co-occur with argumentative support and thereby act as a strong indicator of inference? Might antithesis act as a weak contra-indicator of conflict? The final step is to substantiate or repudiate these hypotheses: (Lawrence et al., 2017c) presents an initial investigation in this direction, though any one of the types of rhetorical figure could be interestingly challenging to identify and highly correlated with some aspect of argumentative structure.

11.2.2 Speaker Profiling[5]

The increasing availability of argumentatively annotated text corpora of appropriate size and quality opens up new possibilities for applying quantitative empirical methods in the study of argumentation. In particular, the use of corpus-based metrics to model the rhetorical profile of a speaker; characterising their style of arguing in terms of their selection of argument schemes, the type of standpoints advanced, and the speech acts used. Such a profile could then be used to help inform argument mining: knowing that a speaker frequently uses a particular argumentation scheme increases the likelihood that it will be found in their utterances; a preference for factual premises may equally make supporting arguments easier to identify.

As a preliminary study, the US2016G1tv corpus was analysed, comparing the styles on the speakers. During the debate, Donald Trump introduced 455 argumentative relations (of inference, conflict and rephrase), while Hilary Clinton accumulated a much lower total of 235. As expected in political debates, both Clinton and Trump regularly made use of Arguments from Example, Cause to Effect, Sign, and Consequences. Striking is Trump's propensity for personal attacks: 15% of his arguments consist of Circumstantial/Generic Ad

[5]The material in this section was presented at the 2019 European Conference on Argumentation (`http://ecargument.org`) as part of the presentation: "Quantitative rhetorical profiling".

Hominem or Argument from Bias, compared to 7% of Clinton's. Trump also uses a considerably higher number of Fear Appeals to justify his standpoints: 10 for Trump (making up 3,4% of his total number of arguments), against 1 for Clinton (0,5%). Clinton, on the other hand, relies more heavily on Popular Opinion/Practice argument schemes than Trump does: 10 counts for Clinton (constituting 5% of her arguments) against 4 counts for Trump (1%). Furthermore, she employs the Argument from Values 10 times (5,2% of her arguments), while Trump only relies on values 5 times (1,7%).

Another stark difference in rhetorical choices made by Clinton and Trump is the type of claims defended. In 28% of the cases, Clinton argues for some policy proposal. In comparison, only 9% of Trump's arguments defend policy proposals. This distinctive difference in rhetorical style is further confirmed by the candidates' use of the Practical Reasoning argument scheme, in which a plan of action is defended on the basis of a particular goal: 17% of Clinton's arguments constitute Practical Reasoning, against 4% of Trump's.

Finally, in terms of speech acts, Trump restates or paraphrases notably more than Clinton does. Trump introduces 112 rephrase relations, while Clinton only uses rephrase 19 times (8%).

Highlighting such differences between speakers offers great potential for analysing their arguments in future material. Further investigation will be required in order to determine how transferable these attributes are, with some likely to depend on the context and other participants, whilst some may reflect more general style.

11.2.3 Argumentation Schemes

Recent work in the field of argumentation schemes is opening up new possibilities for the automatic identification of scheme instances, and the improvement of related argument mining techniques.

(Lawrence et al., 2019a) presents an online annotation assistant combining a novel annotation method for Walton's typology of schemes (Walton et al., 2008), with the widely used OVA software for argument analysis. This annotation method is referred to as the Argument Scheme Key (ASK) (Visser

et al., 2021; Lawrence et al., 2019b). The ASK is a dichotomous identification key that leads the analyst through a series of binary choices based on the distinctive features of subsets of argument schemes until they reach a particular scheme label. The choices are informed by grouping together scheme types in Walton's taxonomy that share particular characteristics. For example, the ASK starts by distinguishing between source-based and other arguments. Each subsequent choice in the key leads to either a particular argument scheme, or to a further choice. For example, an analyst goes through the following sequence of characteristics in identifying an argument as an instance of argument from popular opinion: source-based; about the source's opinion; based on existing opinion; source is a group of people.

These advances in scheme annotation offer the promise of faster and more accurate annotation of scheme data (Visser et al., 2018c). More interestingly still, they also constitute an intermediate step in the development of automated classifiers, utilising the uniquely identifying characteristics of the ASK, with the answer to simpler questions, such as "is the source a group of people or an individual", likely being easier to determine automatically, compared to full scheme instance identification.

11.2.4 Intertextual Argument Mining

The majority of existing argument mining techniques are confined to the particular source text in which an opinion is expressed. The task of *Intertextual Argument Mining*, is that of identifying argumentative relations between topically related texts from different areas. Preliminary work has looked at determining such connections between the US presidential election debates, and corresponding reactions from online discussion (Visser et al., 2018b).

Intertextual Argument Mining shares much in common with the comment-to-article linking task (Aker et al., 2015) which aims to connect readers' comments to the news article segments which they refer to. Indeed, more recent work in comment-to-article linking (Aker et al., 2016) has extended this to include consideration of sentiment and argument structure, assigning a label of *agree*, *disagree* or *neutral* to each article segment-comment pair.

As with comment-to-article linking, preliminary work on Intertextual Argument Mining has mainly employed similarity features, though several additional aspects are available in the case of linking televised debates to their live online reaction, including: explicit references made to a speaker in the debate; temporal ordering of comments (a comment cannot refer to something that hadn't been said at the time it was posted); and, the context of a comment (clusters of comments may all be referencing the same, or similar, points in the debate). Current results show that combining these features gives an accuracy of 0.36 compared to an accuracy of 0.57 for testing with human annotators.

This is a challenging task with characteristics that preclude the use of many techniques that have proven successful in previous argument mining work. However, it is hoped that, by linking argumentative structures together in this way, large interconnected datasets can be created, and the vision of the integrated (World Wide) Argument Web (Rahwan et al., 2007) brought to fruition.

11.3 Concluding remarks

Significant progress has been made in the field of argument mining since the publication of the first paper dedicated to the subject in 2007 (Moens et al., 2007), and high-profile success stories such as IBM's Project Debater[6] continue to push innovation and spark interest in the area. This progress has, in large part, been driven by parallel advances in computational linguistics, artificial intelligence, and machine learning; with the more basic statistical classifiers of a decade ago, slowly giving way to the ever improving results offered by end-to-end and neural network based approaches (Eger et al., 2017; Persing and Ng, 2016; Shnarch et al., 2018; Niven and Kao, 2019).

However argument mining remains an enormously challenging task; as Moens (2018) points out, "a lot of content is not expressed explicitly but resides in the mind of communicator and audience". It is perhaps in this aspect, of understanding the implicit intentions of the communicator and the

[6] `https://www.research.ibm.com/artificial-intelligence/project-debater/`

corresponding interpretation made by the audience, that the greatest future progress in the field can be realised. Many arguments may be signalled, not by explicit linguistic cues, but by the more subtle usage of an argumentation scheme, or the juxtaposition of two topics where one is a frequent source of supporting arguments for the other.

Such subtleties of communication have long been at the core of philosophical research in the analysis, modelling and understanding of argumentation. It is from the blending of these theoretical aspects of argumentation along with the application of computational linguistic techniques, that the techniques presented in this thesis are derived. By virtue of this approach, the work presented here offers two principal contributions: the development of a range of argument mining techniques grounded in argumentation theory; and, the introduction of *Explainable Argument Mining* (XAM).

The individual techniques presented have been shown to produce robust results on their own, but, much as the theoretical works on which they are based each illuminate different and complementary aspects of human understanding, it is in their combination that a more full and accurate picture emerges.

In combination the techniques presented here have been shown to produce comparable results to state of the art techniques developed specifically for use on the datasets tested, whilst maintaining explainability, and working across genres without modification. However, this is still a starting point for XAM, and, as the new techniques and improvements discussed in Section 11.2 are developed, these results are expected to improve further still.

Bibliography

Abbott, R., Ecker, B., Anand, P., and Walker, M. A. (2016). Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), pages 4445–4452, Portoroz, Slovenia.

Accuosto, P. and Saggion, H. (2019). Transferring knowledge from discourse to arguments: A case study with scientific abstracts. In Proceedings of the 6th Workshop on Argument Mining, pages 41–51, Florence, Italy. Association for Computational Linguistics.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access, 6:52138–52160.

Addawood, A. and Bashir, M. (2016). What is your evidence? A study of controversial topics in social media. In Proceedings of the 3rd Workshop on Argumentation Mining, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Agarwal, S. and Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. Bioinformatics, 25(23):3174–3180.

Aharoni, E., Polnarov, A., Lavee, T., Hershcovich, D., Levy, R., Rinott, R., Gutfreund, D., and Slonim, N. (2014). A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In Proceedings of the First Workshop on Argumentation Mining, pages 64–68, Baltimore, MD. Association for Computational Linguistics.

Ailomaa, M. and Rajman, M. (2009). Enhancing natural language search in meeting data with visual meeting overviews. In Proceedings of the 10th Annual Conference of the NZ ACM Special Interest Group on Human-Computer Interaction (CHINZ 2009), pages 6–7, Auckland, New Zealand.

Ajjour, Y., Chen, W.-F., Kiesel, J., Wachsmuth, H., and Stein, B. (2017). Unit segmentation of argumentative texts. In Proceedings of the 4th Workshop on Argument Mining, pages 118–128, Copenhagen, Denmark. Association for Computational Linguistics.

Aker, A., Celli, F., Funk, A., Kurtic, E., Hepple, M., and Gaizauskas, R. (2016). Sheffield-trento system for sentiment and argument structure enhanced comment-to-article linking in the online news domain.

Aker, A., Kurtić, E., Hepple, M., Gaizauskas, R., and Di Fabbrizio, G. (2015). Comment-to-article linking in the online news domain. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 245–249.

Aker, A., Sliwa, A., Ma, Y., Lui, R., Borad, N., Ziyaei, S., and Ghobadi, M. (2017). What works and what does not: Classifier and feature analysis for argument mining. In Proceedings of the 4th Workshop on Argument Mining, pages 91–96, Copenhagen, Denmark. Association for Computational Linguistics.

Al Khatib, K., Wachsmuth, H., Lang, K., Herpel, J., Hagen, M., and Stein, B. (2018). Modeling deliberative argumentation strategies on wikipedia. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2545–2555, Melbourne, Australia. Association for Computational Linguistics.

Al-Khatib, K., Wachsmuth, H., Stein, Hagen, M., Köhler, J., and Stein, B. (2016). Cross-domain mining of argumentative text through distant supervision. In Proceedings of NAACL-HLT, pages 1395–1404, San Diego, CA.

Anand, P., King, J., Boyd-Graber, J., Wagner, E., Martell, C., Oard, D., and Resnik, P. (2011). Believe me–we can do this! annotating persuasive acts in blog text. In Proceedings of the 11th International Workshop on Computational Models of Natural Argument (CMNA 2011) at AAAI 2011, pages 11–15, San Francisco, CA.

Aristotle (1958). Topics. Oxford University Press.

Aristotle (1991). On Rhetoric. Oxford University Press.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In HLT-SS '11 Proceedings of the ACL 2011 Student Session, pages 81–87, Portland, OR.

Atkinson, K., Bench-Capon, T., and Mcburney, P. (2005). A dialogue game protocol for multi-agent argument over proposals for action. Autonomous Agents and Multi-Agent Systems, 11(2):153–171.

Austin, J. L. (1962). How to do things with words. Harvard University Press, 2nd edition.

Awadallah, R., Ramanath, M., and Weikum, G. (2012). Harmony and dissonance: Organizing the people's voices on political controversies. In Proceedings of the fifth ACM international conference on Web search and data mining, pages 523–532, Seattle, WA. ACM.

Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., and Slonim, N. (2017). Stance classification of context-dependent claims. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, volume 1, pages 251–261, Valencia, Spain.

Bar-Haim, R., Krieger, D., Toledo-Ronen, O., Edelstein, L., Bilu, Y., Halfon, A., Katz, Y., Menczel, A., Aharonov, R., and Slonim, N. (2019). From surrogacy to adoption; from bitcoin to cryptocurrency: Debate topic expansion. In Proceedings of the 57th Annual Meeting of the Association for

Computational Linguistics, pages 977–990, Florence, Italy. Association for Computational Linguistics.

Barker, E. and Gaizauskas, R. (2016). Summarizing multi-party argumentative conversations in reader comment on news. In Proceedings of the 3rd Workshop on Argumentation Mining, pages 12–20, Berlin, Germany. Association for Computational Linguistics.

Bex, F., Gordon, T. F., Lawrence, J., and Reed, C. (2012). Interchanging arguments between Carneades and AIF – Theory and practice. In Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012), pages 390–397, Vienna, Austria. IOS Press.

Bex, F., Lawrence, J., and Reed, C. (2014a). Generalising argument dialogue with the dialogue game execution platform. In Parsons, S., Oren, N., Reed, C., and Cerutti, F., editors, Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014), pages 141–152, Pitlochry, Scotland. IOS Press.

Bex, F., Lawrence, J., Snaith, M., and Reed, C. (2013). Implementing the argument web. Communications of the ACM, 56(10):66–73.

Bex, F., Snaith, M., Lawrence, J., and Reed, C. (2014b). Argublogging: An application for the argument web. Web Semantics: Science, Services and Agents on the World Wide Web, 25:9–15.

Bilu, Y., Gera, A., Hershcovich, D., Sznajder, B., Lahav, D., Moshkowich, G., Malet, A., Gavron, A., and Slonim, N. (2019). Argument invention from first principles. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1013–1026, Florence, Italy. Association for Computational Linguistics.

Bilu, Y., Hershcovich, D., and Slonim, N. (2015). Automatic claim negation: Why, how and when. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 84–93, Denver, CO. Association for Computational Linguistics.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022.

Boltužić, F. and Šnajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In Proceedings of the First Workshop on Argumentation Mining, pages 49–58, Baltimore, MD. Association for Computational Linguistics.

Boltužić, F. and Šnajder, J. (2015). Identifying prominent arguments in online debates using semantic textual similarity. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 110–115, Denver, CO. Association for Computational Linguistics.

Bosc, T., Cabrio, E., and Villata, S. (2016). DART: a dataset of arguments and their relations on twitter. In Proceedings of the 10th edition of the Language Resources and Evaluation Conference, pages 1258–1263, Portoroz, Slovenia.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7):107–117.

Brown, D. W. (2017). Clinton-trump corpus. Retrieved from `http://www.thegrammarlab.com`.

Budzynska, K. (2011). Araucaria-PL: Software for teaching argumentation theory. In Proceedings of the Third International Congress on Tools for Teaching Logic (TICTTL 2011), pages 30–37, Salamanca, Spain.

Budzynska, K., Janier, M., Reed, C., and Saint-Dizier, P. (2016). Theoretical foundations for illocutionary structure parsing. Argument & Computation, 7(1):91–108.

Budzynska, K., Janier, M., Reed, C., Saint-Dizier, P., Stede, M., and Yaskorska, O. (2014). A model for processing illocutionary structures and argumentation in debates. In Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC), pages 917–924, Reykjavik, Iceland.

Budzynska, K. and Reed, C. (2011). Whence inference. Technical report, University of Dundee.

Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A. C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., et al. (2010). Towards an iso standard for dialogue act annotation. In Prceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta.

Buschmann, T. and Bystrykh, L. V. (2013). Levenshtein error-correcting barcodes for multiplexed dna sequencing. BMC bioinformatics, 14(1):272.

Cabrio, E., Tonelli, S., and Villata, S. (2013). From discourse analysis to argumentation schemes and back: Relations and differences. In International Workshop on Computational Logic in Multi-Agent Systems, pages 1–17. Springer.

Cabrio, E. and Villata, S. (2012). Generating abstract arguments: a natural language approach. In Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012), pages 454–461, Vienna, Austria. IOS Press.

Cabrio, E. and Villata, S. (2018). Five years of argument mining: a data-driven analysis. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 5427–5433, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. Computational linguistics, 22(2):249–254.

Carlile, W., Gurrapadi, N., Ke, Z., and Ng, V. (2018). Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia. Association for Computational Linguistics.

Carstens, L. and Toni, F. (2015). Towards relation based argumentation mining. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 29–34, Denver, CO. Association for Computational Linguistics.

Carstens, L., Toni, F., and Evripidou, V. (2014). Argument mining and social debates. In Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014), pages 451–452, Pitlochry, Scotland. IOS Press.

Chesñevar, C., Modgil, S., Rahwan, I., Reed, C., Simari, G., South, M., Vreeswijk, G., Willmott, S., et al. (2006). Towards an argument interchange format. The Knowledge Engineering Review, 21(04):293–316.

Choi, Y., Jung, Y., and Myaeng, S.-H. (2010). Identifying controversial issues and their sub-topics in news articles. In Chen, H., Chau, M., Li, S., Urs, S., Srinivasa, S., and Wang, G. A., editors, Intelligence and Security Informatics, pages 140–153. Springer.

Cialdini, R. B. (2001). Influence: Science and practice, volume 4. Allyn and Bacon Boston, MA.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3):273–297.

Dagan, I., Glickman, O., and Magnini, B. (2006). The PASCAL recognising textual entailment challenge. In Quinonero-Candela, J., Dagan, I., Magnini, B., and d'Alche Buc, F., editors, Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment, pages 177–190. Springer.

Dahlgren, K. (1988). Naive semantics for natural language understanding. Springer.

Das, S. and Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance

Association Annual Conference (APFA), volume 35, pages 1–16, Bangkok, Thailand.

Daxenberger, J., Eger, S., Habernal, I., Stab, C., and Gurevych, I. (2017). What is the essence of a claim? cross-domain claim identification. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

De Marneffe, M.-C. and Manning, C. D. (2008). The Stanford typed dependencies representation. In COLING 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, pages 1–8, Manchester, United Kingdom. Association for Computational Linguistics.

Delmonte, R. (2007). Computational Linguistic Text Processing: Logical Form, Semantic Interpretation, Discourse Relations and Question Answering. Nova Publishers.

Dori-Hacohen, S. and Allan, J. (2013). Detecting controversy on the web. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pages 1845–1848, San Francisco, CA. ACM.

Dusmanu, M., Cabrio, E., and Villata, S. (2017). Argument mining on twitter: Arguments, facts and sources. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Duthie, R., Budzynska, K., and Reed, C. (2016a). Mining ethos in political debate. In Baroni, P., Stede, M., and Gordon, T., editors, Proceedings of the Sixth International Conference on Computational Models of Argument (COMMA 2016), pages 299–310, Potsdam, Germany. IOS Press.

Duthie, R., Lawrence, J., Budzynska, K., and Reed, C. (2016b). The CASS technique for evaluating the performance of argument mining. In Proceedings of the 3rd Workshop on Argumentation Mining, pages 40–49, Berlin, Germany. Association for Computational Linguistics.

Eckle-Kohler, J., Kluge, R., and Gurevych, I. (2015). On the role of discourse markers for discriminating claims and premises in argumentative discourse. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2236–2242.

Egawa, R., Morio, G., and Fujita, K. (2019). Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 422–428, Florence, Italy.

Eger, S., Daxenberger, J., and Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. In Proceedings of ACL, pages 11–22, Vancouver, Canada.

Fahnestock, J. and Secor, M. (1988). The stases in scientific and literary argument. Written Communication, 5(4):427–443.

Feng, V. W. and Hirst, G. (2011). Classifying arguments by scheme. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 987–996, Portland, OR. Association for Computational Linguistics.

Feng, V. W. and Hirst, G. (2014). Two-pass discourse segmentation with pairing and global features. CoRR, abs/1407.8215.

Firoozeh, N., Nazarenko, A., Alizon, F., and Daille, B. (2020). Keyword extraction: Issues and methods. Natural Language Engineering, 26(3):259–291.

Freeman, J. B. (1991). Dialectics and the macrostructure of arguments: A theory of argument structure, volume 10. Walter de Gruyter.

Freeman, J. B. (2000). What types of statements are there? Argumentation, 14(2):135–157.

Freeman, J. B. (2011). Argument Structure: Representation and Theory. Springer.

Galassi, A., Lippi, M., and Torroni, P. (2018). Argumentative link prediction using residual networks and multi-objective learning. In Proceedings of the 5th Workshop on Argument Mining, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

Gao, X., Xiao, B., Tao, D., and Li, X. (2010). A survey of graph edit distance. Pattern Analysis and applications, 13(1):113–129.

Gemechu, D. and Reed, C. (2019). Decompositional argument mining: A general purpose approach for argument graph construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 516–526, Florence, Italy. Association for Computational Linguistics.

Ghosh, D., Muresan, S., Wacholder, N., Aakhus, M., and Mitsui, M. (2014). Analyzing argumentative discourse units in online interactions. In Proceedings of the First Workshop on Argumentation Mining, pages 39–48, Baltimore, MD. Association for Computational Linguistics.

Gipp, B. and Beel, J. (2010). Citation based plagiarism detection: a new approach to identify plagiarized work language independently. In Proceedings of the 21st ACM conference on Hypertext and hypermedia, pages 273–274. ACM.

Givón, T. (1983). Topic continuity in discourse: A quantitative cross-language study, volume 3. John Benjamins Publishing.

Gleize, M., Shnarch, E., Choshen, L., Dankin, L., Moshkowich, G., Aharonov, R., and Slonim, N. (2019). Are you convinced? choosing the more convincing evidence with a Siamese network. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Gordon, T. F., Prakken, H., and Walton, D. (2007). The carneades model of argument and burden of proof. Artificial Intelligence, 171(10):875–896.

Goudas, T., Louizos, C., Petasis, G., and Karkaletsis, V. (2014). Argument

extraction from news, blogs, and social media. In Artificial Intelligence: Methods and Applications, pages 287–299. Springer.

Green, N. (2014). Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In Proceedings of the First Workshop on Argumentation Mining, pages 11–18, Baltimore, MD. Association for Computational Linguistics.

Green, N. (2015). Identifying argumentation schemes in genetics research articles. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 12–21, Denver, CO. Association for Computational Linguistics.

Green, N. (2018a). Proposed method for annotation of scientific arguments in terms of semantic relations and argument schemes. In Proceedings of the 5th Workshop on Argument Mining, Brussels, Belgium. Association for Computational Linguistics.

Green, N. L. (2018b). Towards mining scientific discourse using argumentation schemes. Argument & Computation, 9(2):121–135.

Grennan, W. (1997). Informal Logic: Issues and Techniques. McGill-Queen's Press-MQUP.

Grimes, J. E. (1975). The thread of discourse, volume 207. Walter de Gruyter.

Groarke, L., Tindale, C., and Fisher, L. (1997). Good reasoning matters! : a constructive approach to critical thinking. Oxford University Press, Toronto.

Grosse, K., Chesñevar, C. I., and Maguitman, A. G. (2012). An argument-based approach to mining opinions from twitter. In First International Conference on Agreement Technologies (AT 2012), pages 408–422, Dubrovnik, Croatia. Citeseer.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. Computational linguistics, 12(3):175–204.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stump, S., and Yang, G.-Z. (2019). XAI - Explainable artificial intelligence. Science Robotics, 4(37).

Habernal, I. and Gurevych, I. (2015). Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2127–2137, Lisbon, Portugal.

Habernal, I. and Gurevych, I. (2017). Argumentation mining in user-generated web discourse. Computational Linguistics, 43(1):125–179.

Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B. (2018). SemEval-2018 task 12: The argument reasoning comprehension task. In Proceedings of the 12th International Workshop on Semantic Evaluation, pages 763–772, New Orleans, LA. Association for Computational Linguistics.

Harrell, M. (2005). Using argument diagramming software in the classroom. Teaching Philosophy, 28(2):163–177.

Hassan, N., Li, C., and Tremayne, M. (2015). Detecting check-worthy factual claims in presidential debates. In Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM '15, pages 1835–1838, Melbourne, Australia. ACM.

Hastings, A. C. (1963). A Reformulation of the Modes of Reasoning in Argumentation. PhD thesis, Northwestern University.

Hidey, C. and McKeown, K. (2018). Persuasive influence detection: The role of argument sequencing. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA.

Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. Computational linguistics, 19(3):501–530.

Hitchcock, D. (1985). Enthymematic arguments. Informal Logic, 7(2):289–98.

Hobbs, J. R. (1993). Intention, information, and structure in discourse: A first draft. In Burning Issues in Discourse, NATO Advanced Research Workshop, pages 41–66, Maratea, Italy.

Hoeken, H. and Hustinx, L. (2003). The relative persuasiveness of different types of evidence. In van Eemeren, F. H., Blair, J. A., Willard, C. A., and Snoeck Henkemans, A. F., editors, Proceedings of the fifth conference of the International Society for the Study of Argumentation, pages 497–501, Amsterdam, Netherlands. SicSat.

Hogenboom, A., Hogenboom, F., Kaymak, U., Wouters, P., and De Jong, F. (2010). Mining economic sentiment using argumentation structures. In et al., T. J., editor, Advances in Conceptual Modeling–Applications and Challenges, pages 200–209. Springer.

Holmes, G., Donkin, A., and Witten, I. H. (1994). Weka: A machine learning workbench. In Proceedings of the 1994 second Australian and New Zealand Conference on Intelligent Information Systems, pages 357–361, Brisbane, Australia. IEEE.

Hou, Y. and Jochim, C. (2017). Argument relation classification using a joint inference model. In Proceedings of the 4th Workshop on Argument Mining, pages 60–66, Copenhagen, Denmark. Association for Computational Linguistics.

Houngbo, H. and Mercer, R. (2014). An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In Proceedings of the First Workshop on Argumentation Mining, pages 19–23, Baltimore, MD. Association for Computational Linguistics.

Hua, X. and Wang, L. (2017). Neural argument generation augmented with externally retrieved evidence. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 219–230, Vancouver, Canada. Association for Computational Linguistics.

Hubert, L. (1977). Kappa revisited. Psychological Bulletin, 84(2):289–297.

Hutchinson, B. (2004). Acquiring the meaning of discourse markers. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pages 684–691, Barcelona, Spain. Association for Computational Linguistics.

Janier, M., Lawrence, J., and Reed, C. (2014). OVA+: An argument analysis interface. In Parsons, S., Oren, N., Reed, C., and Cerutti, F., editors, Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014), pages 463–464, Pitlochry, Scotland. IOS Press.

Janier, M. and Reed, C. (2016). Corpus resources for dispute mediation discourse. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), pages 1014–1021, Portoroz, Slovenia.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The ICSI meeting corpus. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), volume 1, pages 364–367, Hong Kong, China. IEEE.

Jaradat, I., Gencheva, P., Barrón-Cedeño, A., Màrquez, L., and Nakov, P. (2018). Claimrank: Detecting check-worthy claims in arabic and english. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 26–30, New Orleans, LA. Association for Computational Linguistics.

Jia, L., Yu, C., and Meng, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In Proceedings of the 18th ACM conference on Information and knowledge management, pages 1827–1830.

Jo, Y., Visser, J., Reed, C., and Hovy, E. (2019). A cascade model for proposition extraction in argumentation. In Proceedings of the 6th Workshop on Argument Mining, pages 11–24, Florence, Italy. Association for Computational Linguistics.

Katzav, J. and Reed, C. (2004). On argumentation schemes and the natural classification of arguments. Argumentation, 18(2):239–259.

Ke, Z., Carlile, W., Gurrapadi, N., and Ng, V. (2018). Learning to give feedback: Modeling attributes affecting argument persuasiveness in student

essays. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 4130–4136, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Kendall, M. G. (1938). A new measure of rank correlation. Biometrika, 30(1/2):81–93.

Kienpointner, M. (1992). Alltagslogik: struktur und funktion von argumentationsmustern. Frommann-Holzboog.

Kiesel, J., Al Khatib, K., Hagen, M., and Stein, B. (2015). A shared task on argumentation mining in newspaper editorials. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 35–38, Denver, CO. Association for Computational Linguistics.

Kim, S.-M. and Hovy, E. (2006a). Automatic identification of pro and con reasons in online reviews. In Proceedings of COLING/ACL 2006, pages 483–490, Sydney, Australia. Association for Computational Linguistics.

Kim, S.-M. and Hovy, E. (2006b). Extracting opinions, opinion holders, and topics expressed in online news media text. In Proceedings of the Workshop on Sentiment and Subjectivity in Text, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Kirschner, C., Eckle-Kohler, J., and Gurevych, I. (2015). Linking the thoughts: Analysis of argumentation structures in scientific publications. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 1–11, Denver, CO. Association for Computational Linguistics.

Kirschner, P. A., Buckingham-Shum, S. J., and Carr, C. S. (2003). Visualizing argumentation: Software tools for collaborative and educational sense-making. Springer.

Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. (2007). He says, she says: Conflict and coordination in wikipedia. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 453–462, San Jose, CA. ACM.

Knott, A. (1996). A data-driven methodology for motivating a set of coherence relations. PhD thesis, Department of Artificial Intelligence, University of Edinburgh.

Konat, B., Lawrence, J., Park, J., Budzynska, K., and Reed, C. (2016). A corpus of argument networks: Using graph properties to analyse divisive issues. In Proceedings of the 10th edition of the Language Resources and Evaluation Conference.

Krauthoff, T., Meter, C., Betz, G., Baurmann, M., and Mauve, M. (2018). D-BAS: A Dialog-Based Online Argumentation System. In Computational Models of Argument (COMMA), pages 325–336, Warsaw, Poland.

Krippendorff, K. (1980). Content analysis: An introduction to its methodology. Sage publications.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 3:159–174.

Lasnik, H. and Uriagereka, J. (1988). A course in GB syntax: Lectures on binding and empty categories. MIT Press, Cambridge, MA.

Lauscher, A., Glavaš, G., and Eckert, K. (2018a). Arguminsci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In Proceedings of the 5th Workshop on Argument Mining, pages 22–28, Brussels, Belgium. Association for Computational Linguistics.

Lauscher, A., Glavaš, G., and Ponzetto, S. P. (2018b). An argument-annotated corpus of scientific publications. In Proceedings of the 5th Workshop on Argument Mining, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

Lavee, T., Orbach, M., Kotlerman, L., Kantor, Y., Gretz, S., Dankin, L., Jacovi, M., Bilu, Y., Aharonov, R., and Slonim, N. (2019). Towards effective rebuttal: Listening comprehension using corpus-wide claim mining. In Proceedings of the 6th Workshop on Argument Mining, pages 58–66, Florence, Italy. Association for Computational Linguistics.

Lawrence, J., Bex, F., and Reed, C. (2012a). Dialogues on the argument web: Mixed initiative argumentation with arvina. In Proceedings of the 4th International Conference on Computational Models of Argument (COMMA 2012), pages 513–514, Vienna, Austria. IOS Press.

Lawrence, J., Bex, F., Reed, C., and Snaith, M. (2012b). AIFdb: Infrastructure for the argument web. In Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012), pages 515–516, Vienna, Austria. IOS Press.

Lawrence, J., Duthie, R., Budzysnka, K., and Reed, C. (2016). Argument analytics. In Baroni, P., Stede, M., and Gordon, T., editors, Proceedings of the Sixth International Conference on Computational Models of Argument (COMMA 2016), pages 371–378, Berlin. IOS Press.

Lawrence, J., Janier, M., and Reed, C. (2015). Working with open argument corpora. In Proceedings of the 1st European Conference on Argumentation (ECA 2015), Lisbon. College Publications.

Lawrence, J., Park, J., Budzynska, K., Cardie, C., Konat, B., and Reed, C. (2017a). Using argumentative structure to interpret debates in online deliberative democracy and erulemaking. ACM Transactions on Internet Technology (TOIT), 17(3):25.

Lawrence, J. and Reed, C. (2014). AIFdb corpora. In Parsons, S., Oren, N., Reed, C., and Cerutti, F., editors, Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014), pages 465–466, Pitlochry, Scotland. IOS Press.

Lawrence, J. and Reed, C. (2015). Combining argument mining techniques. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 127–136, Denver, CO. Association for Computational Linguistics.

Lawrence, J. and Reed, C. (2016). Argument mining using argumentation scheme structures. In Baroni, P., Stede, M., and Gordon, T., editors, Proceedings of the Sixth International Conference on Computational Models

of Argument (COMMA 2016), pages 379–390, Potsdam, Germany. IOS Press.

Lawrence, J. and Reed, C. (2017a). Mining argumentative structure from natural language text using automatically generated premise-conclusion topic models. In Proceedings of the 4th Workshop on Argument Mining, pages 39–48, Copenhagen, Denmark. Association for Computational Linguistics.

Lawrence, J. and Reed, C. (2017b). Using complex argumentative interactions to reconstruct the argumentative structure of large-scale debates. In Proceedings of the Fourth Workshop on Argumentation Mining, Copenhagen. Association for Computational Linguistics.

Lawrence, J. and Reed, C. (2020). Argument mining: A survey. Computational Linguistics, 45(4):765–818.

Lawrence, J., Reed, C., Allen, C., McAlister, S., and Ravenscroft, A. (2014). Mining arguments from 19th century philosophical texts using topic based modelling. In Proceedings of the First Workshop on Argumentation Mining, pages 79–87, Baltimore, MD. Association for Computational Linguistics.

Lawrence, J., Snaith, M., Konat, B., Budzynska, K., and Reed, C. (2017b). Debating technology for dialogical argument: Sensemaking, engagement, and analytics. ACM Transactions on Internet Technology (TOIT), 17(3):24:1–24:23.

Lawrence, J., Visser, J., and Reed, C. (2017c). Harnessing rhetorical figures for argument mining. Argument & Computation, 8(3):289–310.

Lawrence, J., Visser, J., and Reed, C. (2018). BBC Moral Maze: Test your argument. In Modgil, S., Budzynska, K., and Lawrence, J., editors, Proceedings of the Seventh International Conference on Computational Models of Argument (COMMA 2018), pages 465–466, Warsaw. IOS Press.

Lawrence, J., Visser, J., and Reed, C. (2019a). An online annotation assistant for argument schemes. In Proceedings of the 13th Linguistic Annotation

Workshop, pages 100–107, Florence, Italy. Association for Computational Linguistics.

Lawrence, J., Visser, J., Walton, D., and Reed, C. (2019b). A decision tree for annotating argumentation scheme corpora. In 3rd European Conference on Argumentation (ECA 2019), pages 97–114, Groningen, Netherlands.

Le, D. T., Nguyen, C.-T., and Nguyen, K. A. (2018). Dave the debater: a retrieval-based and generative argumentative dialogue agent. In Proceedings of the 5th Workshop on Argument Mining, pages 121–130, Brussels, Belgium. Association for Computational Linguistics.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In ICML, volume 14, pages 1188–1196.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8):707–710.

Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., and Slonim, N. (2014). Context dependent claim detection. In Proceedings of the 25th International Conference on Computational Linguistics, pages 1489–1500, Dublin, Ireland.

Levy, R., Gretz, S., Sznajder, B., Hummel, S., Aharonov, R., and Slonim, N. (2017). Unsupervised corpus–wide claim detection. In Proceedings of the 4th Workshop on Argument Mining, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In Machine learning: ECML-98, pages 4–15. Springer.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, page 605. Association for Computational Linguistics.

Lindahl, A., Borin, L., and Rouces, J. (2019). Towards assessing argumentation annotation - a first step. In Proceedings of the 6th Workshop on Argument Mining, pages 177–186, Florence, Italy. Association for Computational Linguistics.

Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2:627–666.

Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. In Proceedings of the Third IEEE International Conference on Data Mining (ICDM-03), pages 179–186. IEEE.

Madnani, N., Heilman, M., Tetreault, J., and Chodorow, M. (2012). Identifying high-level organizational elements in argumentative discourse. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 20–28, Montreal, Canada. Association for Computational Linguistics.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: A theory of text organization. University of Southern California, Information Sciences Institute.

Marwell, G. and Schmitt, D. R. (1967). Dimensions of compliance-gaining behavior: An empirical analysis. Sociometry, 30(4):350–364.

Mayer, T., Cabrio, E., and Villata, S. (2018). Evidence type classification in randomized controlled trials. In Proceedings of the 5th Workshop on Argument Mining, pages 29–34, Brussels, Belgium. Association for Computational Linguistics.

McBurney, P. and Parsons, S. (2009). Dialogue games for agent argumentation. In Simari, G. and Rahwan, I., editors, Argumentation in artificial intelligence, pages 261–280. Springer.

Merity, S., Murphy, T., and Curran, J. R. (2009). Accurate argumentative zoning with maximum entropy models. In Proceedings of the 2009 Workshop

on Text and Citation Analysis for Scholarly Digital Libraries, pages 19–26, Suntec City, Singapore. Association for Computational Linguistics.

Metzinger, T. (1999). Teaching philosophy with argumentation maps review of can computers think? the debate by Robert E. Horn. PSYCHE, 5.

Mihalcea, R., Corley, C., Strapparava, C., et al. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence, volume 6, pages 775–780.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.

Miller, G. A. (1995). Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

Misra, A., Ecker, B., and Walker, M. (2016). Measuring the similarity of sentential arguments in dialogue. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 276–287, Los Angeles. Association for Computational Linguistics.

Mochales, R. and Ieven, A. (2009). Creating an argumentation corpus: do theories apply to real arguments?: a case study on the legal argumentation of the echr. In Proceedings of the 12th International Conference on Artificial Intelligence and Law, pages 21–30, Barcelona, Spain. ACM.

Mochales, R. and Moens, M.-F. (2011). Argumentation mining. Artificial Intelligence and Law, 19:1–22.

Moens, M.-F. (2018). Argumentation mining: How can a machine acquire common sense and world knowledge? Argument & Computation, 9(1):1–14.

Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007). Automatic detection of arguments in legal texts. In Proceedings of the 11th international

conference on Artificial intelligence and law, pages 225–230, Stanford, CA. ACM.

Morio, G. and Fujita, K. (2018). End-to-end argument mining for discussion threads based on parallel constrained pointer architecture. In Proceedings of the 5th Workshop on Argument Mining, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Murdock, J., Allen, C., Borner, K., Light, R., McAlister, S., Ravenscroft, A., Rose, R., Rose, D., Otsuka, J., Bourget, D., Lawrence, J., and Reed, C. (2017). Multi-level computational methods for interdisciplinary research in the hathitrust digital library. PLOS ONE, 12(9):1–21.

Musi, E., Ghosh, D., and Muresan, S. (2016). Towards feasible guidelines for the annotation of argument schemes. In Proceedings of the 3rd Workshop on Argumentation Mining, pages 82–93, Berlin, Germany. Association for Computational Linguistics.

Naderi, N. and Hirst, G. (2018a). Automated fact-checking of claims in argumentative parliamentary debates. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 60–65, Brussels, Belgium. Association for Computational Linguistics.

Naderi, N. and Hirst, G. (2018b). Using context to identify the language of face-saving. In Proceedings of the 5th Workshop on Argument Mining, pages 111–120, Brussels, Belgium. Association for Computational Linguistics.

Newell, E., Margolin, D., and Ruths, D. (2018). An attribution relations corpus for political news. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 100–108. Association for Computational Linguistics.

Nguyen, H. and Litman, D. (2015). Extracting argument and domain words for identifying argument components in texts. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 22–28, Denver, CO. Association for Computational Linguistics.

Nguyen, H. V. and Litman, D. (2016). Context-aware argumentative relation mining. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany. Association for Computational Linguistics.

Niculae, V. (2018). Learning Deep Models with Linguistically-Inspired Structure. PhD thesis, Cornell University.

Niculae, V., Park, J., and Cardie, C. (2017). Argument mining with structured svms and rnns. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Niven, T. and Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Okada, A., Shum, S. J. B., and Sherborne, T. (2008). Knowledge Cartography: software tools and mapping techniques. Springer.

Ong, N., Litman, D., and Brusilovsky, A. (2014). Ontology-based argument mining and automatic essay scoring. In Proceedings of the First Workshop on Argumentation Mining, pages 24–28, Baltimore, MD. Association for Computational Linguistics.

Opitz, J. and Frank, A. (2019). Dissecting content and context in argumentative relation analysis. In Proceedings of the 6th Workshop on Argument Mining, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In Proceedings of the 12th international conference on artificial intelligence and law, pages 98–107, Barcelona, Spain. ACM.

Pallotta, V. and Delmonte, R. (2011). Automatic argumentative analysis for interaction mining. Argument & Computation, 2(2-3):77–106.

Pallotta, V., Ghorbel, H., Ballim, A., Lisowska, A., and Marchand-Maillet, S. (2004). Towards meeting information systems. In Proceedings of the 6th International Conference in Enterprise Information Systems (ICEIS), pages 464–469, Porto,Pourtugal.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2:1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the ACL 2002 conference on Empirical methods in natural language processing (EMNLP 2002), pages 79–86, Pennsylvania, PA. Association for Computational Linguistics.

Pantel, P. A. (2003). Clustering by committee. PhD thesis, University of Alberta.

Park, J. and Cardie, C. (2014). Identifying appropriate support for propositions in online user comments. In Proceedings of the First Workshop on Argumentation Mining, pages 29–38, Baltimore, MD. Association for Computational Linguistics.

Park, J. and Cardie, C. (2018). A corpus of erulemaking user comments for measuring evaluability of arguments. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pages 1623–1628, Miyazaki, Japan. European Language Resources Association (ELRA).

Passon, M., Lippi, M., Serra, G., and Tasso, C. (2018). Predicting the usefulness of Amazon reviews using off-the-shelf argumentation mining. In Proceedings of the 5th Workshop on Argument Mining, pages 35–39, Brussels, Belgium. Association for Computational Linguistics.

Patwari, A., Goldwasser, D., and Bagchi, S. (2017). Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In Proceedings of the 2017 ACM Conference on Information and Knowledge Management, CIKM '17, pages 2259–2262, Singapore. ACM.

Pease, A., Budzynska, K., Lawrence, J., and Reed, C. (2014). Lakatos games for mathematical argument. In Parsons, S., Oren, N., Reed, C., and Cerutti, F., editors, Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014), pages 59–66, Pitlochry, Scotland. IOS Press.

Pease, A., Lawrence, J., Budzynska, K., Corneli, J., and Reed, C. (2017). Lakatos-style collaborative mathematics through dialectical, structured and abstract argumentation. Artificial Intelligence, 246:181–219.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. J. Mach. Learn. Res., 12:2825–2830.

Peldszus, A. (2014). Towards segment-based recognition of argumentation structure in short texts. In Proceedings of the First Workshop on Argumentation Mining, pages 88–97, Baltimore, MD. Association for Computational Linguistics.

Peldszus, A. (2018). Automatic recognition of argumentation structure in short monological texts. PhD thesis, Department of Linguistics, University of Potsdam.

Peldszus, A. and Stede, M. (2013a). From argument diagrams to argumentation mining in texts: A survey. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 7(1):1–31.

Peldszus, A. and Stede, M. (2013b). Ranking the annotators: An agreement study on argumentation structure. In Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, pages 196–204, Sofia, Bulgaria.

Peldszus, A. and Stede, M. (2015). Towards detecting counter-considerations in text. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 104–109, Denver, CO. Association for Computational Linguistics.

Peldszus, A. and Stede, M. (2016). An annotated corpus of argumentative microtexts. In Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2, pages 801–815, London. College Publications.

Perelman, C. and Olbrechts-Tyteca, L. (1969). The New Rhetoric: A Treatise on Argumentation. University of Notre Dame Press.

Persing, I. and Ng, V. (2015). Modeling argument strength in student essays. In Proceedings of ACL, pages 543–552, Beijing, China.

Persing, I. and Ng, V. (2016). End-to-end argumentation mining in student essays. In Proceedings of NAACL-HLT, pages 1384–1394, San Diego, CA.

Petasis, G. and Karkaletsis, V. (2016). Identifying argument components through textrank. In Proceedings of the 3rd Workshop on Argumentation Mining, pages 94–102, Berlin, Germany. Association for Computational Linguistics.

Piao, S., Ananiadou, S., Tsuruoka, Y., Sasaki, Y., and McNaught, J. (2007). Mining opinion polarity relations of citations. In International Workshop on Computational Semantics (IWCS), pages 366–371, Tilburg, Netherlands.

Polanyi, L. (1988). A formal model of the structure of discourse. Journal of pragmatics, 12(5):601–638.

Pollock, J. (1986). Contemporary Theories of Knowledge. Rowman And Littlefield, Towota, NJ.

Pollock, J. L. (1987). Defeasible reasoning. Cognitive science, 11(4):481–518.

Pollock, J. L. (1995). Cognitive carpentry: A blueprint for how to build a person. MIT Press.

Potash, P., Romanov, A., and Rumshisky, A. (2017). Here's my point: Joint pointer architecture for argument mining. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1364–1373, Copenhagen, Denmark. Association for Computational Linguistics.

Prakken, H. (2005). Coherence and flexibility in dialogue games for argumentation. Journal of logic and computation, 15(6):1009–1040.

Rahwan, I., Zablith, F., and Reed, C. (2007). Laying the foundations for a world wide argument web. Artificial Intelligence, 171:897–921.

Razuvayevskaya, O. and Teufel, S. (2017). Finding enthymemes in real-world texts: A feasibility study. Argument & Computation, 8(2):113–129.

Reed, C. (2006). Preliminary results from an argument corpus. In Bermúdez, E. M. and Miyares, L. R., editors, Linguistics in the twenty-first century, pages 185–196. Cambridge Scholars Press.

Reed, C., Budzynska, K., Duthie, R., Janier, M., Konat, B., Lawrence, J., Pease, A., and Snaith, M. (2017). The argument web: an online ecosystem of tools, systems and services for argumentation. Philosophy & Technology, 30(2):137–160.

Reed, C., Mochales Palau, R., Rowe, G., and Moens, M.-F. (2008). Language resources for studying argument. In Proceedings of the 6th Language Resources and Evaluation Conference (LREC-2008), pages 91–100, Marrakech, Morocco.

Reed, C. and Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. International Journal on Artificial Intelligence Tools, 13(4):961–980.

Reynolds, R. A. and Reynolds, J. L. (2002). Evidence. In Dillard, J. P. and Pfau, M., editors, The persuasion handbook: Developments in theory and practice, pages 427–444. Sage, Thousand Oaks, CA.

Rienks, R., Heylen, D., and Weijden, v. d. E. (2005). Argument diagramming of meeting conversations. In Multimodal Multiparty Meeting Processing, Workshop at the 7th Intl. Conference on Multimodal Interfaces, pages 85–92, Trento, Italy. IOS Press.

Rigotti, E. and Morasso, S. G. (2010). Comparing the argumentum model of topics to other contemporary approaches to argument schemes: the procedural and material components. Argumentation, 24(4):489–512.

Rinott, R., Dankin, L., Perez, C. A., Khapra, M. M., Aharoni, E., and Slonim, N. (2015). Show me your evidence-an automatic method for context dependent evidence detection. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 440–450, Lisbon, Portugal.

Robertson, D. (2004). A lightweight coordination calculus for agent systems. In International Workshop on Declarative Agent Languages and Technologies, pages 183–197, New York, NY. Springer.

Rumshisky, A., Gronas, M., Potash, P., Dubov, M., Romanov, A., Kulshreshtha, S., and Gribov, A. (2017). Combining network and language indicators for tracking conflict intensity. In International Conference on Social Informatics, pages 391–404, Oxford, United Kingdom. Springer.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 50(4):696–735.

Saint-Dizier, P. (2012). Processing natural language arguments with the <TextCoop> platform. Argument & Computation, 3(1):49–82.

Saint-Dizier, P. (2017). Knowledge-driven argument mining based on the qualia structure. Argument & Computation, 8(2):193–210.

Saint-Dizier, P. (2018). A two-level approach to generate synthetic argumentation reports. Argument & Computation, 9(2):137–154.

Sardianos, C., Katakis, I. M., Petasis, G., and Karkaletsis, V. (2015). Argument extraction from news. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 56–66, Denver, CO. Association for Computational Linguistics.

Scheuer, O., Loll, F., Pinkwart, N., and McLaren, B. M. (2010). Computer-supported argumentation: A review of the state of the art. International Journal of Computer-Supported Collaborative Learning, 5(1):43–102.

Schulz, C., Eger, S., Daxenberger, J., Kahse, T., and Gurevych, I. (2018). Multi-task learning for argumentation mining in low-resource settings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 35–41, New Orleans, LA. Association for Computational Linguistics.

Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. Cambridge university press.

Shnarch, E., Alzate, C., Dankin, L., Gleize, M., Hou, Y., Choshen, L., Aharonov, R., and Slonim, N. (2018). Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Skeppstedt, M., Peldszus, A., and Stede, M. (2018). More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In Proceedings of the 5th Workshop on Argument Mining, pages 155–163, Brussels, Belgium. Association for Computational Linguistics.

Snaith, M., Lawrence, J., and Reed, C. (2010). Mixed initiative argument in public deliberation. In F., D. C., A., M., and C., P., editors, From e-Participation to Online Deliberation, Proceedings of the Fourth International Conference on Online Deliberation (OD2010), pages 2–13, Leeds, UK.

Snaith, M., Medellin, R., Lawrence, J., and Reed, C. (2017). Arguers and the Argument Web. In Bex., F., Grasso, F., Green, N., Paglieri, F., and Reed, C., editors, Argument Technologies: Theory, Analysis & Applications, pages 57–72. College Publications.

Sobhani, P., Inkpen, D., and Matwin, S. (2015). From argumentation mining to stance classification. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 67–77, Denver, CO. Association for Computational Linguistics.

Soricut, R. and Marcu, D. (2003). Sentence-level discourse parsing using syntactic and lexical information. In Proc. of the Human Language Technology Conference of the North American Chapter of the ACL, pages 149–156, Edmonton, Canada.

Sparck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation.

Stab, C. and Gurevych, I. (2014a). Annotating argument components and relations in persuasive essays. In Proceedings of the 25th International Conference on Computational Linguistics, pages 1501–1510, Dublin, Ireland.

Stab, C. and Gurevych, I. (2014b). Identifying argumentative discourse structures in persuasive essays. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Stab, C. and Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. Computational Linguistics, 43(3):619–659.

Stede, M. and Schneider, J. (2018). Argumentation Mining. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., and Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In Proceedings of the 25th international conference on world wide web, pages 613–624, Montreal, Canada. International World Wide Web Conferences Steering Committee.

Teufel, S., Carletta, J., and Moens, M.-F. (1999). An annotation scheme for discourse-level argumentation in research articles. In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1493–1502, Singapore. Association for Computational Linguistics.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006). Automatic classification of citation function. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Toulmin, S. E. (1958). The uses of argument. Cambridge University Press.

Traum, D. (2017). Computational approaches to dialogue. In Weigand, E., editor, The Routledge Handbook of Language and Dialogue, pages 143–161. Taylor & Francis.

van Eemeren, F. H., Garssen, B., Krabbe, E. C., Snoeck Henkemans, A., Verheij, B., and Wagemans, J. H. (2014). Handbook of Argumentation Theory. Springer.

van Eemeren, F. H., Houtlosser, P., and Snoeck Henkemans, A. F. (2007). Argumentative Indicators in Discourse: A Pragma-Dialectical Study. Argumentation Library. Springer.

van Gelder, T. (2007). The rationale for rationale. Law, probability and risk, 6(1-4):23–42.

Van Lent, M., Fisher, W., and Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the Sixteenth Conference on Innovative Applications of Artificial Intelligence, pages 900–907. San Jose, California, USA.

van Rijsbergen, C. J. (1979). Information Retrieval. Butterworth.

Villalba, M. P. G. and Saint-Dizier, P. (2012). Some facets of argument mining for opinion analysis. In Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012), pages 23–34, Vienna, Austria. IOS Press.

Visser, J., Duthie, R., Lawrence, J., and Reed, C. (2018a). Intertextual correspondence for integrating corpora. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC), pages 3511–3517, Miyazaki, Japan. European Language Resources Association (ELRA).

Visser, J., Duthie, R., Lawrence, J., and Reed, C. (2018b). Intertextual Correspondence for Integrating Corpora. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pages 3511–3517, Miyazaki, Japan. European Language Resources Association (ELRA).

Visser, J., Konat, B., Duthie, R., Koszowy, M., Budzynska, K., and Reed, C. (2020a). Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. Language Resources and Evaluation, 54(1):123–154.

Visser, J., Lawrence, J., and Reed, C. (2020b). Reason-checking fake news. Communications of the ACM, 63(11):38–40.

Visser, J., Lawrence, J., Reed, C., Wagemans, J., and Walton, D. (2021). Annotating argument schemes. Argumentation, 35:101–139.

Visser, J., Lawrence, J., Wagemans, J., and Reed, C. (2018c). Revisiting computational models of argument schemes: Classification, annotation, comparison. In Modgil, S., Budzynska, K., and Lawrence, J., editors, Proceedings of the Seventh International Conference on Computational Models of Argument (COMMA 2018), pages 313–324, Warsaw. IOS Press.

Wachsmuth, H., Potthast, M., Al-Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., and Stein, B. (2017a). Building an argument search engine for the web. In Proceedings of the 4th Workshop on Argument Mining, pages 49–59, Copenhagen, Denmark. Association for Computational Linguistics.

Wachsmuth, H., Stein, B., and Ajjour, Y. (2017b). "PageRank" for argument relevance. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, volume 1, pages 1117–1127, Valencia, Spain.

Wachsmuth, H., Syed, S., and Stein, B. (2018). Retrieval of the best counterargument without prior topic knowledge. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 241–251, Melbourne, Australia.

Walker, M. A., Tree, J. E. F., Anand, P., Abbott, R., and King, J. (2012). A corpus for research on deliberation and debate. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), pages 812–817, Istanbul, Turkey.

Walker, V., Vazirova, K., and Sanford, C. (2014). Annotating patterns of reasoning about medical theories of causation in vaccine cases: Toward a type system for arguments. In Proceedings of the First Workshop on

Argumentation Mining, pages 1–10, Baltimore, MD. Association for Computational Linguistics.

Walker, V. R., Foerster, D., Ponce, J. M., and Rosen, M. (2018). Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment. In Proceedings of the 5th Workshop on Argument Mining, pages 68–78, Brussels, Belgium. Association for Computational Linguistics.

Walton, D. (1996). Argumentation schemes for presumptive reasoning. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Walton, D. (2011). Argument mining by applying argumentation schemes. Studies in Logic, 4(1):38–64.

Walton, D., Reed, C., and Macagno, F. (2008). Argumentation Schemes. Cambridge University Press.

Webber, B., Egg, M., and Kordoni, V. (2011). Discourse structure and language technology. Natural Language Engineering, 18(4):437–490.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing, pages 347–354, Vancouver, Canada. Association for Computational Linguistics.

Wyner, A., Peters, W., and Price, D. (2015). Argument discovery and extraction with the argument workbench. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 78–83, Denver, CO. Association for Computational Linguistics.

Wyner, A., Schneider, J., Atkinson, K., and Bench-Capon, T. (2012). Semi-automated argumentative analysis of online product reviews. In Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012), pages 43–50, Vienna, Austria. IOS Press.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 32, pages 5753–5763. Curran Associates, Inc.